Bernoulli **19**(3), 2013, 846–885
DOI: 10.3150/12-BEJ468

Divergence rates of Markov order estimators and their application to statistical estimation of stationary ergodic processes

ZSOLT TALATA

Department of Mathematics, University of Kansas, 405 Snow Hall, 1460 Jayhawk Boulevard, Lawrence, KS 66045-7523, USA. E-mail: talata@math.ku.edu

Stationary ergodic processes with finite alphabets are estimated by finite memory processes from a sample, an n-length realization of the process, where the memory depth of the estimator process is also estimated from the sample using penalized maximum likelihood (PML). Under some assumptions on the continuity rate and the assumption of non-nullness, a rate of convergence in $\bar{d}$-distance is obtained, with explicit constants. The result requires an analysis of the divergence of PML Markov order estimators for not necessarily finite memory processes. This divergence problem is investigated in more generality for three information criteria: the Bayesian information criterion with generalized penalty term yielding the PML, and the normalized maximum likelihood and the Krichevsky–Trofimov code lengths. Lower and upper bounds on the estimated order are obtained. The notion of consistent Markov order estimation is generalized for infinite memory processes using the concept of oracle order estimates, and generalized consistency of the PML Markov order estimator is presented.

Keywords: finite memory estimator; infinite memory; information criteria; Markov approximation; minimum description length; oracle inequalities; penalized maximum likelihood; rate of convergence

1. Introduction

This paper is concerned with the problem of estimating stationary ergodic processes with finite alphabet from a sample, an observed length n realization of the process, with the $\bar{d}$-distance being considered between the process and the estimated one. The $\bar{d}$-distance was introduced by Ornstein [26] and became one of the most widely used metrics over stationary processes. Two stationary processes are close in $\bar{d}$-distance if there is a joint distribution whose marginals are the distributions of the processes such that the marginal processes are close with high probability (see Section 5 for the formal definition). The class of ergodic processes is $\bar{d}$-closed and entropy is $\bar{d}$-continuous, which properties do not hold for the weak topology [34].

Ornstein and Weiss [27] proved that for stationary processes isomorphic to i.i.d. processes, the empirical distribution of the $k(n)$-length blocks is a strongly consistent estimator of the $k(n)$-length parts of the process in $\bar{d}$-distance if and only if $k(n) \leq (\log n)/h$, where h denotes the entropy of the process.

Csiszár and Talata [13] estimated the n-length part of a stationary ergodic process X by a Markov process of order k_n. The transition probabilities of this Markov estimator process are the empirical conditional probabilities, and the order $k_n \to +\infty$ does not depend on the sample. They obtained a rate of convergence of the Markov estimator to the process X in $\bar{d}$-distance, which consists of two terms. The first one is the bias due to the error of the approximation of the process by a Markov chain. The second term is the variation due to the error of the estimation of the parameters of the Markov chain from a sample.

In this paper, the order k_n of the Markov estimator process is estimated from the sample. For the order estimation, penalized maximum likelihood (PML) with general penalty term is used. The resulted Markov estimator process finds a tradeoff between the bias and the variation as it uses shorter memory for faster memory decays of the process X. If the process X is a Markov chain, the PML order estimation recovers its order asymptotically with a wide range of penalty terms.

Not only an asymptotic rate of convergence result is obtained but also an explicit bound on the probability that the $\bar{d}$-distance of the above Markov estimator from the process X is greater than ε. It is assumed that the process X is non-null, that is, the conditional probabilities of the symbols given the pasts are separated from zero, and that the continuity rate of the process X is summable and the restricted continuity rate is uniformly convergent. These conditions are usually assumed in this area [6, 17, 18, 25]. The summability of the continuity rate implies that the process is isomorphic to an i.i.d. process [4].

The above result on statistical estimation of stationary ergodic processes requires a non-asymptotic analysis of the Markov order estimation for not necessarily finite memory processes. In this paper, this problem is also investigated in more generality: under milder conditions than it would be needed for the above bound and not only for the PML method.

A popular approach to the Markov order estimation is the minimum description length (MDL) principle [3, 28]. This method evaluates an information criterion for each candidate order based on the sample and the estimator takes the order for which the value is minimal. The normalized maximum likelihood (NML) [36] and the Krichevsky–Trofimov (KT) [23] code lengths are natural information criteria because the former minimizes the worst case maximum redundancy for the model class of k-order Markov chains, while the latter does so, up to an additive constant, with the average redundancy. The Bayesian information criterion (BIC) [33] can be regarded as an approximation of the NML and KT code lengths. The PML is a generalization of BIC; special settings of the penalty term yield the BIC and other well-known information criteria, such as the Akaike information criterion (AIC) [1]. There are other methods for Markov order estimation, see [19] and references there, and the problem can also be formulated in the setting of hypothesis testing [29].

If a process is a Markov chain, the NML and KT Markov order estimators are strongly consistent if the candidate orders have an upper bound $o(\log n)$ [9]. Without such a bound, they fail to be consistent [11]. The BIC Markov order estimator is strongly consistent without any bound on the candidate orders [11]. If a process has infinite memory, the Markov order estimators are expected to tend to infinity as $n \to +\infty$. The concept of context trees of arbitrary stationary ergodic processes is a model more complex than Markov chains. Recent results [12] in that area imply that this expectation holds true for the BIC and KT Markov order estimators but they provide no information about the asymptotics of the divergence.

In this paper, the divergence of the PML, NML and KT Markov order estimators for not necessarily finite memory processes is investigated. Not only asymptotic rates of divergence are obtained but also explicit bounds on the probability that the estimators are greater and less, respectively, than some order. Instead of the usual assumption of non-nullness, it is assumed only that the conditional probabilities of one of the symbols given the pasts are separated from zero. This property is called weakly non-nullness and is "noticeably weaker" than non-nullness [7].

First, the process is assumed to be weakly non-null and α-summable. The α-summability [14, 15, 21, 24] is a condition weaker than the summability of the continuity rate. Under these conditions, a bound on the probability that the estimators are greater than some order is obtained, that yields an $\mathcal{O}(\log n)$ upper bound on the estimated order eventually almost surely as $n \to +\infty$.

Then, a bound on the probability that the estimators are less than some order is obtained assuming that the process is weakly non-null and the decay of its continuity rates is in some exponential range. This bound implies that the estimators satisfying the conditions attain a $c \log n$ divergence rate eventually almost surely as $n \to +\infty$, where the coefficient c depends on the range of the continuity rates. The class of processes with exponentially decaying continuity rate is considered in various problems [17, 20]. Fast divergence rate of the estimators are expected only for a certain range of continuity rates. Clearly, the estimators do not have a fast divergence rate if the memory decay of the process is too fast. On the other hand, too slow memory decay is also not favored to a fast divergence rate because then the empirical probabilities do not necessarily converge to the true probabilities.

To provide additional insight into the asymptotics of Markov order estimators, the notion of consistent Markov order estimation is generalized for infinite memory processes. A Markov order estimator is compared to its oracle version, which is calculated based on the true distribution of the process instead of the empirical distribution. The oracle concept is used in various problems, see, for example, [2, 5, 16, 22]. If the decay of the continuity rate of the process is faster than exponential, the ratio of the PML Markov order estimator with sufficiently large penalty term to its oracle version is shown to converge to 1 in probability.

The structure of the paper is the following. In Section 2, notation and definitions are introduced for stationary ergodic processes with finite alphabets. In Section 3, the PML, NML and KT information criteria are introduced. Section 4 contains the results on divergence of the information-criterion based Markov order estimators. In Section 5,

the problem of estimating stationary ergodic process in $\bar{d}$-distance is formulated and our results are presented. The results require bounds on empirical entropies, which are stated in Section 4 and are proved in Section 6. Section 7 contains the proof of the divergence results, and Section 8 the proof of the process estimation results.

2. Finite and infinite memory processes

Let $X = \{X_i, -\infty < i < +\infty\}$ be a stationary ergodic stochastic process with finite alphabet A. We write $X_i^j = X_i, \ldots, X_j$ and $x_i^j = x_i, \ldots, x_j \in A^{j-i+1}$ for $j \geq i$. If $j < i$, x_i^j is the empty string. For two strings $x_1^i \in A^i$ and $y_1^j \in A^j$, $x_1^i y_1^j$ denotes their concatenation $x_1, \ldots, x_i, y_1, \ldots, y_j \in A^{i+j}$. Write

$$P(x_i^j) = \Pr(X_i^j = x_i^j)$$

and, if $P(x_{-m}^{-1}) > 0$,

$$P(a|x_{-m}^{-1}) = \Pr(X_0 = a | X_{-m}^{-1} = x_{-m}^{-1}).$$

For $m = 0$, $P(a|x_{-m}^{-1}) = P(a)$.

The process X is called *weakly non-null* if

$$\alpha_0 = \sum_{a \in A} \inf_{x_{-\infty}^{-1} \in A^\infty} P(a|x_{-\infty}^{-1}) > 0.$$

Letting

$$\alpha_k = \min_{y_{-k}^{-1} \in A^k} \sum_{a \in A} \inf_{x_{-\infty}^{-1} \in A^\infty : x_{-k}^{-1} = y_{-k}^{-1}} P(a|x_{-\infty}^{-1}), \qquad k = 1, 2, \ldots,$$

we say that the process X is *α-summable* if

$$\alpha = \sum_{k=0}^{+\infty} (1 - \alpha_k) < +\infty.$$

The *continuity rates* of the process X are

$$\bar{\gamma}(k) = \sup_{x_{-\infty}^{-1} \in A^\infty} \sum_{a \in A} |P(a|x_{-k}^{-1}) - P(a|x_{-\infty}^{-1})|$$

and

$$\underline{\gamma}(k) = \inf_{x_{-\infty}^{-1} \in A^\infty} \sum_{a \in A} |P(a|x_{-k}^{-1}) - P(a|x_{-\infty}^{-1})|.$$

Obviously, $\underline{\gamma}(k) \leq \bar{\gamma}(k)$. If $\sum_{k=1}^{\infty} \bar{\gamma}(k) < +\infty$, then the process X is said to have *summable continuity rate*.

Remark 2.1. Since for any $x_{-k}^{-1} \in A^k$ and $z_{-m}^{-k-1} \in A^{m-k}$, $m \geq k$,

$$\inf_{x_{-\infty}^{-k-1}} P(a|x_{-\infty}^{-1}) \leq P(a|z_{-m}^{-k-1}x_{-k}^{-1}) \leq \sup_{x_{-\infty}^{-k-1}} P(a|x_{-\infty}^{-1}),$$

the above definition of continuity rate is equivalent to

$$\bar{\gamma}(k) = \sup_{i>k} \max_{x_{-i}^{-1} \in A^i} \sum_{a \in A} |P(a|x_{-k}^{-1}) - P(a|x_{-i}^{-1})|.$$

Remark 2.2. The process is α-summable if it has summable continuity rate because

$$\begin{aligned} 1 - \alpha_k &\leq 1 - \max_{y_{-k}^{-1} \in A^k} \sum_{a \in A} P(a|y_{-k}^{-1}) \\ &\quad + \max_{y_{-k}^{-1} \in A^k} \sum_{a \in A} \sup_{x_{-\infty}^{-1} \in A^\infty : x_{-k}^{-1} = y_{-k}^{-1}} (P(a|y_{-k}^{-1}) - P(a|x_{-\infty}^{-1})) \\ &\leq |A| \bar{\gamma}(k). \end{aligned}$$

The k-order *entropy* of the process X is

$$H_k = -\sum_{a_1^k \in A^k} P(a_1^k) \log P(a_1^k), \qquad k \geq 1,$$

and the k-order *conditional entropy* is

$$h_k = -\sum_{a_1^{k+1} \in A^{k+1}} P(a_1^{k+1}) \log P(a_{k+1}|a_1^k), \qquad k \geq 0.$$

Logarithms are to the base 2. It is well known for stationary processes [8, 10] that the conditional entropy h_k is a non-negative decreasing function of k, therefore its limit exists as $k \to +\infty$. The *entropy rate* of the process is

$$\bar{H} = \lim_{k \to +\infty} h_k = \lim_{k \to +\infty} \frac{1}{k} H_k.$$

Note that $h_k - \bar{H} \geq 0$ for any $k \geq 0$.

The process X is a *Markov chain* of order k if for each $n > k$ and $x_1^n \in A^n$

$$P(x_1^n) = P(x_1^k) \prod_{i=k+1}^{n} P(x_i|x_{i-k}^{i-1}), \tag{2.1}$$

where $P(x_1^k)$ is called initial distribution and $\{P(a|a_1^k), a \in A, a_1^k \in A^k\}$ is called transition probability matrix. The case $k = 0$ corresponds to i.i.d. processes. The process X is of *infinite memory* if it is not a Markov chain for any order $k < +\infty$. For infinite memory processes, $h_k - \bar{H} > 0$ for any $k \geq 0$.

In this paper, we consider statistical estimates based on a sample X_1^n, an n-length part of the process. Let $N_n(a_1^k)$ denote the number of occurrences of the string a_1^k in the sample X_1^n

$$N_n(a_1^k) = |\{i : X_{i+1}^{i+k} = a_1^k, 0 \le i \le n-k\}|.$$

For $k \ge 1$, the empirical probability of the string a_1^k is

$$\hat{P}(a_1^k) = \frac{N_n(a_1^k)}{n-k+1}$$

and the empirical conditional probability of $a_{k+1} \in A$ given a_1^k is

$$\hat{P}(a_{k+1}|a_1^k) = \frac{N_n(a_1^{k+1})}{N_{n-1}(a_1^k)}.$$

For $k = 0$, $\hat{P}(a_{k+1}|a_1^k) = \hat{P}(a_{k+1})$. The k-order *empirical entropy* is

$$\hat{H}_k(X_1^n) = -\sum_{a_1^k \in A^k} \hat{P}(a_1^k) \log \hat{P}(a_1^k), \qquad 1 \le k \le n,$$

and the k-order *empirical conditional entropy* is

$$\hat{h}_k(X_1^n) = -\sum_{a_1^{k+1} \in A^{k+1}} \hat{P}(a_1^{k+1}) \log \hat{P}(a_{k+1}|a_1^k), \qquad 0 \le k \le n-1.$$

The likelihood of the sample X_1^n with respect to a k-order Markov chain model of the process X with some transition probability matrix $\{Q(a_{k+1}|a_1^k), a_{k+1} \in A, a_1^k \in A^k\}$, by (2.1), is

$$P'(X_1^n) = P'(X_1^k) \prod_{a_1^{k+1} \in A^{k+1}} Q(a_{k+1}|a_1^k)^{N_n(a_1^{k+1})}.$$

For $0 \le k < n$, the *maximum likelihood* is the maximum in $Q(a_{k+1}|a_1^k)$ of the second factor above, which equals

$$\mathrm{ML}_k(X_1^n) = \prod_{a_1^{k+1} \in A^{k+1}} \hat{P}(a_{k+1}|a_1^k)^{N_n(a_1^{k+1})}.$$

Note that $\log \mathrm{ML}_k(X_1^n) = -(n-k)\hat{h}_k(X_1^n)$.

3. Information criteria

An information criterion assigns a score to each hypothetical model (here, Markov chain order) based on a sample, and the estimator will be that model whose score is minimal.

Definition 3.1. *For an information criterion*

$$\mathrm{IC}_{X_1^n}(\cdot) : \mathbb{N} \to \mathbb{R}^+,$$

the Markov order estimator is

$$\hat{k}_{\mathrm{IC}}(X_1^n) = \arg \min_{0 \le k < n} \mathrm{IC}_{X_1^n}(k).$$

Remark 3.2. Here, the number of candidate Markov chain orders based on a sample is finite, therefore the minimum is attained. If the minimizer is not unique, the smallest one will be taken as arg min.

We consider three, the most frequently used information criteria, namely, the Bayesian information criterion and its generalization, the family of penalized maximum likelihood (PML) [11, 33], the normalized maximum likelihood (NML) code length [36], and the Krichevsky–Trofimov (KT) code length [23].

Definition 3.3. *Given a penalty function* $\mathrm{pen}(n)$*, a non-decreasing function of the sample size* n*, for a candidate order* $0 \le k < n$ *the PML criterion is*

$$\begin{aligned} \mathrm{PML}_{X_1^n}(k) &= -\log \mathrm{ML}_k(X_1^n) + (|A|-1)|A|^k \,\mathrm{pen}(n) \\ &= (n-k)\hat{h}_k(X_1^n) + (|A|-1)|A|^k \,\mathrm{pen}(n). \end{aligned}$$

The k-order Markov chain model of the process X is described by the conditional probabilities $\{Q(a_{k+1}|a_1^k), a_{k+1} \in A, a_1^k \in A^k\}$, and $(|A|-1)|A|^k$ of these are free parameters.

The second term of the PML criterion, which is proportional to the number of free parameters of the k-order Markov chain model, is increasing in k. The first term, for a given sample, is known to be decreasing in k. Hence, minimizing the criterion yields a tradeoff between the goodness of fit of the sample to the model and the complexity of the model.

Remark 3.4. If $\mathrm{pen}(n) = \frac{1}{2}\log n$, the PML criterion is called *Bayesian information criterion* (BIC), and if $\mathrm{pen}(n) = 1$, *Akaike information criterion* (AIC).

The minimum description length (MDL) principle minimizes the length of a code of the sample tailored to the model class. Strictly speaking, the information criterion would have an additive term, the length of a code of the structure parameter. This additional term, the length of a code of k, is omitted since it does not affect the results.

Definition 3.5. *For a candidate order* $0 \le k < n$*, the NML criterion is*

$$\mathrm{NML}_{X_1^n}(k) = -\log P_{\mathrm{NML},k}(X_1^n),$$

where

$$P_{\mathrm{NML},k}(X_1^n) = \frac{\mathrm{ML}_k(X_1^n)}{\Sigma(n,k)} \qquad \text{with } \Sigma(n,k) = \sum_{x_1^n \in A^n} \mathrm{ML}_k(x_1^n)$$

is the k-order NML-probability of X_1^n.

Remark 3.6. Writing

$$\mathrm{NML}_{X_1^n}(k) = -\log \mathrm{ML}_k(X_1^n) + \log \Sigma_{n,k},$$

the NML criterion can be regarded as a PML criterion in a broader sense.

Definition 3.7. *For a candidate order $0 \le k < n$, the KT criterion is*

$$\mathrm{KT}_{X_1^n}(k) = -\log P_{\mathrm{KT},k}(X_1^n),$$

where

$$P_{\mathrm{KT},k}(X_1^n) = \frac{1}{|A|^k} \prod_{\substack{a_1^k \in A^k: \\ N_{n-1}(a_1^k) \ge 1}} \frac{\prod_{a_{k+1}: N_n(a_1^{k+1}) \ge 1} [(N_n(a_1^{k+1}) - 1/2)(N_n(a_1^{k+1}) - 3/2) \cdots (1/2)]}{(N_{n-1}(a_1^k) - 1 + |A|/2)(N_{n-1}(a_1^k) - 2 + |A|/2) \cdots (|A|/2)}$$

is the k-order KT-probability of X_1^n. (For $k = 0$, $N_{n-1}(a_1^k) = n$.)

Remark 3.8. The k-order KT-probability of the sample is equal to a mixture of the probabilities of the sample with respect to all k-order Markov chains with uniform initial distribution, where the mixture distribution over the transition probability matrices $\{Q(a_{k+1}|a_1^k), a_{k+1} \in A, a_1^k \in A^k\}$ is independent for the rows $Q(\cdot|a_1^k)$, $a_1^k \in A^k$, and has Dirichlet $(\frac{1}{2}, \ldots, \frac{1}{2})$ distribution in the rows. Hence, the KT Markov order estimator can be regarded as a Bayes (maximum a posteriori) estimator.

Remark 3.9. The k-order NML and KT coding distributions are nearly optimal among the k-order Markov chains, in the sense that the code lengths $\lceil -\log P_{\mathrm{NML},k}(X_1^n) \rceil$ and $\lceil -\log P_{\mathrm{KT},k}(X_1^n) \rceil$ minimize the worst case maximum and average, respectively, redundancy for this class (up to an additive constant in the latter case).

4. Divergence of Markov order estimators

The BIC Markov order estimator is strongly consistent [11], that is, if the process is a Markov chain of order k, then $\hat{k}_{\mathrm{BIC}}(X_1^n) = k$ eventually almost surely as $n \to +\infty$. "Eventually almost surely" means that with probability 1, there exists a threshold n_0 (depending on the infinite realization X_1^∞) such that the claim holds for all $n \ge n_0$. Increasing the penalty term, up to cn, where $c > 0$ is a sufficiently small constant, does

not affect the strong consistency. It is not known whether or not the strong consistency holds for smaller penalty terms but it is known that if the candidate orders are upper bounded by $c\log n$, where $c>0$ is a sufficiently small constant, that is, the estimator minimizes the PML over the orders $0\le k\le c\log n$ only, then $\mathrm{pen}(n)=C\log\log n$ still provides the strong consistency, where $C>0$ is a sufficiently large constant [35].

The NML and KT Markov order estimators fail to be strongly consistent because for i.i.d. processes with uniform distribution, they converge to infinity at a rate $\mathcal{O}(\log n)$ [11]. However, if the candidate orders are upper bounded by $\mathrm{o}(\log n)$, the strong consistency holds true [9].

If the process is of infinite memory, the BIC and KT Markov order estimators diverge to infinity [12]. In this section, results on the divergence rate of the PML, NML and KT Markov order estimators are presented. Bounds on the probability that the estimators are greater and less, respectively, than some order are obtained, with explicit constants. The first implies that under mild conditions, the estimators do not exceed the $\mathcal{O}(\log n)$ rate eventually almost surely as $n\to+\infty$. The second bound implies that the rate $\mathcal{O}(\log n)$ is attained eventually almost surely as $n\to+\infty$ for the processes whose continuity rates decay in some exponential range.

At the end of the section, the notion of consistent Markov order estimation is generalized for infinite memory processes. If the continuity rates decay faster than exponential, the PML Markov order estimator is shown to be consistent with the oracle-type order estimate.

The proofs use bounds on the simultaneous convergence of empirical entropies of orders in an increasing set. These bounds are obtained for finite sample sizes n with explicit constants under mild conditions so they are of independent interest and are also presented here.

Theorem 4.1. *For any weakly non-null and α-summable stationary ergodic process, for any $0<\varepsilon<1/2$*

$$\Pr\left(\max_{1\le k\le(\varepsilon\log n)/(4\log|A|)}|\hat{H}_k(X_1^n)-H_k|>\frac{1}{n^{1/2-\varepsilon}}\right)\le\exp\left(-\frac{c_1\varepsilon^3}{\log n}n^{\varepsilon/2}\right)$$

and

$$\Pr\left(\max_{0\le k\le(\varepsilon\log n)/(4\log|A|)}|\hat{h}_k(X_1^n)-h_k|>\frac{1}{n^{1/2-\varepsilon}}\right)\le\exp\left(-\frac{c_2\varepsilon^3}{\log n}n^{\varepsilon/2}\right),$$

where $c_1,c_2>0$ are constants depending only on the distribution of the process.

Proof. The proof including the explicit expression of the constants is in Section 6. □

Remark 4.2. The convergence of $\hat{H}_{k_n}(X_1^n)$ and $\hat{h}_{k_n}(X_1^n)$, $k_n\to\infty$, to the entropy rate $\bar{H}$ of the process could be investigated using Theorem 4.1. However, good estimates of the entropy rate are known from the theory of universal codes. In particular, mixtures of the KT distributions over all possible orders provide universal codes in the class of all

stationary ergodic processes [29–31], therefore the corresponding code length is a suitable estimate of the entropy rate.

An application of the Borel–Cantelli lemma in Theorem 4.1 yields the following asymptotic result.

Corollary 4.3. *For any weakly non-null and α-summable stationary ergodic process, for any $0<\varepsilon<1/2$*

$$|\hat{H}_k(X_1^n)-H_k|\le\frac{1}{n^{1/2-\varepsilon}}\quad\text{and}\quad|\hat{h}_k(X_1^n)-h_k|\le\frac{1}{n^{1/2-\varepsilon}}$$

simultaneously for all $k\le\frac{\varepsilon\log n}{4\log|A|}$, eventually almost surely as $n\to+\infty$.

Remark 4.4. By [20], under much stronger conditions on the process, the convergence rate of $\hat{H}_k(X_1^n)$ and $\hat{h}_k(X_1^n)$ to $\bar{H}$ is $n^{-1/2}$ for some fixed $k=\mathcal{O}(\log n)$. Hence, the rate in Theorem 4.1 cannot be improved significantly.

The first divergence result of the paper is the following.

Theorem 4.5. *For any weakly non-null and α-summable stationary ergodic process there exist $\lambda_1,\lambda_2>0$ depending only on the distribution of the process, such that for the Markov order estimator $\hat{k}_{\mathrm{IC}}(X_1^n)$*

$$\Pr(\hat{k}_{\mathrm{IC}}(X_1^n)>k_n)\le 2^{\lambda_1+2\log n-\lambda_2 k_n}$$

for any sequence k_n, $n\in\mathbb{N}$, where IC is either the PML with arbitrary $\mathrm{pen}(n)$ *or the NML or the KT criterion.*

Proof. The proof including the explicit expression of the constants is in Section 7. □

An application of the Borel–Cantelli lemma in Theorem 4.5 yields the following asymptotic result.

Corollary 4.6. *For any weakly non-null and α-summable stationary ergodic process there exists a constant $C>0$ such that for the Markov order estimator $\hat{k}_{\mathrm{IC}}(X_1^n)$*

$$\hat{k}_{\mathrm{IC}}(X_1^n)\le C\log n$$

eventually almost surely as $n\to+\infty$, where IC is either the PML with arbitrary $\mathrm{pen}(n)$ *or the NML or the KT criterion.*

The second divergence result is the following.

Theorem 4.7. *For any weakly non-null stationary ergodic process with continuity rates* $\bar{\gamma}(k) \le \delta_1 2^{-\zeta_1 k}$ *and* $\underline{\gamma}(k) \ge \delta_2 2^{-\zeta_2 k}$ *for some* $\zeta_1, \zeta_2, \delta_1, \delta_2 > 0$ *(*$\zeta_2 \ge \zeta_1$*), if*

$$\frac{6 \log |A|}{\zeta_1} \le \varepsilon < \frac{1}{2},$$

the Markov order estimator $\hat{k}_{\mathrm{IC}}(X_1^n)$ *satisfies that*

$$\Pr\left(\hat{k}_{\mathrm{IC}}(X_1^n) \le \frac{1}{2\zeta_2} \left(\frac{1}{2} - \varepsilon \right) \log n - c_3 \right) \le \exp\left(-\frac{c_2 \varepsilon^3}{\log n} n^{\varepsilon/2} \right),$$

if $n \ge n_0$*, where IC is either the PML with* $\mathrm{pen}(n) \le \mathcal{O}(\sqrt{n})$ *or the NML or the KT criterion, and* $c_2, n_0 > 0$*,* $c_3 \in \mathbb{R}$ *are constants depending only on the distribution of the process and* $\mathrm{pen}(n)$*.*

Proof. The proof including the explicit expression of the constants is in Section 7. □

An application of the Borel–Cantelli lemma in Theorem 4.7 yields the following asymptotic result.

Corollary 4.8. *For any weakly non-null stationary ergodic process with continuity rates* $\bar{\gamma}(k) \le \delta_1 2^{-\zeta_1 k}$ *and* $\underline{\gamma}(k) \ge \delta_2 2^{-\zeta_2 k}$ *for some* $\zeta_1, \zeta_2, \delta_1, \delta_2 > 0$ *with* $\zeta_2 \ge \zeta_1 > 12 \log |A|$*, the Markov order estimator* $\hat{k}_{\mathrm{IC}}(X_1^n)$ *satisfies that*

$$\hat{k}_{\mathrm{IC}}(X_1^n) \ge C' \log n$$

eventually almost surely as $n \to +\infty$*, where IC is either the PML with* $\mathrm{pen}(n) \le \mathcal{O}(\sqrt{n})$ *or the NML or the KT criterion, and* $C' > 0$ *is a constant depending only on the distribution of the process.*

The section concludes with the consistency result.

Definition 4.9. *For a candidate order* $0 \le k < n$ *the oracle PML criterion is*

$$\mathrm{PML}_{o,n}(k) = (n-k) h_k + (|A| - 1) |A|^k \, \mathrm{pen}(n),$$

and the oracle PML Markov order estimator is

$$k_{\mathrm{PML},n} = \arg \min_{0 \le k < n} \mathrm{PML}_{o,n}(k).$$

Remark 4.10. For Markov chains of order k, $k_{\mathrm{PML},n} = k$ if n is sufficiently large, with any $\mathrm{pen}(n) = \mathrm{o}(n)$.

Theorem 4.11. *For any weakly non-null stationary ergodic process with*

$$\frac{\log \bar{\gamma}(k)}{k} \to -\infty, \qquad k \to \infty,$$

the PML Markov order estimator $\hat{k}_{\mathrm{PML}}(X_1^n)$ *with* $\mathrm{pen}(n) = n^\kappa$, $\frac{1}{2} < \kappa < 1$, *is consistent in the sense that*

$$\frac{\hat{k}_{\mathrm{PML}}(X_1^n)}{k_{\mathrm{PML},n}} \to 1$$

in probability as $n \to +\infty$.

Proof. The proof is in Section 7. □

5. Statistical estimation of processes

In the results of this section, the divergence rate of Markov order estimators will play a central role. The problem of statistical estimation of stationary ergodic processes by finite memory processes is considered, and the following distance is used. The per-letter Hamming distance between two strings x_1^n and y_1^n is

$$d_n(x_1^n, y_1^n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i \neq y_i) \qquad \text{where } \mathbb{I}(a \neq b) = \begin{cases} 1, & \text{if } a \neq b, \\ 0, & \text{if } a = b, \end{cases}$$

and the $\bar{d}$-*distance* between two random sequences X_1^n and Y_1^n is defined by

$$\bar{d}(X_1^n, Y_1^n) = \min_{\mathbb{P}} \mathbb{E}_{\mathbb{P}} d_n(\tilde{X}_1^n, \tilde{Y}_1^n),$$

where the minimum is taken over all the joint distributions $\mathbb{P}$ of $\tilde{X}_1^n$ and $\tilde{Y}_1^n$ whose marginals are equal to the distributions of X_1^n and Y_1^n.

The process X is estimated by a Markov chain of order $k = k_n$ from the sample in the following way.

Definition 5.1. *The* empirical k-order Markov estimator *of a process X based on the sample X_1^n is the stationary Markov chain, denoted by $\hat{X}[k]$, of order k with transition probability matrix $\{\hat{P}(a_{k+1}|a_1^k), a_{k+1} \in A, a_1^k \in A^k\}$. If the initial distribution of a stationary Markov chain with these transition probabilities is not unique, then any of these initial distributions can be taken.*

In the previous section, weakly non-nullness is assumed for the process. In this section the process X is assumed to be *non-null*, that is,

$$p_{\inf} = \min_{a \in A} \inf_{x_{-\infty}^{-1} \in A^\infty} P(a|x_{-\infty}^{-1}) > 0.$$

Remark 5.2. For any non-null stationary ergodic process, $P(a_1^k) \le (1-p_{\inf})^k$ for any $a_1^k \in A^k$. Hence, Theorem 4.5 holds with $\lambda_1 = 0$ and $\lambda_2 = |\log(1-p_{\inf})|$, see the proof of the theorem.

The assumption of non-nullness allows us to use the following quantity instead of $\underline{\gamma}(k)$. The *restricted continuity rate* of the process X is

$$\bar{\gamma}(k|m) = \max_{x_{-m}^{-1} \in A^m} \sum_{a \in A} |P(a|x_{-k}^{-1}) - P(a|x_{-m}^{-1})|, \qquad k < m.$$

Similarly to Remark 2.1, note that the above definition is equivalent to

$$\bar{\gamma}(k|m) = \max_{k < i \le m} \max_{x_{-i}^{-1} \in A^i} \sum_{a \in A} |P(a|x_{-k}^{-1}) - P(a|x_{-i}^{-1})|.$$

Hence, $\lim_{m \to +\infty} \bar{\gamma}(k|m) = \bar{\gamma}(k)$ for any fixed k. We say that the process X has *uniformly convergent restricted continuity rate* with parameters θ_1, θ_2, k_θ if

$$\bar{\gamma}(k)^{\theta_1} \le \bar{\gamma}(k|\lceil \theta_2 k \rceil) \qquad \text{if } k \ge k_\theta, \text{ for some } \theta_1 \ge 1, \theta_2 > 1.$$

The order k of the empirical Markov estimator $\hat{X}[k]$ is estimated from the sample, using the PML criterion. The estimated order needs to be bounded to guarantee an accurate assessment of the memory decay of the process.

Definition 5.3. *For an information criterion IC, the Markov order estimator bounded by $r_n < n$, $r_n \in \mathbb{N}$, is*

$$\hat{k}_{\mathrm{IC}}(X_1^n | r_n) = \arg \min_{0 \le k \le r_n} \mathrm{IC}_{X_1^n}(k).$$

The optimal order can be smaller than the upper bound if the memory decay of the process is sufficiently fast. Define

$$K_n(r_n, \bar{\gamma}, f(n)) = \min\{\lfloor r_n \rfloor, k \ge 0 : \bar{\gamma}(k) < f(n)\},$$

where $f(n) \searrow 0$ and $r_n \nearrow \infty$. Since $\bar{\gamma}$ is a decreasing function, K_n increases in n but does not exceed r_n. It is less than r_n if $\bar{\gamma}$ vanishes sufficiently fast, and then the faster $\bar{\gamma}$ vanishes, the slower K_n increases.

The process estimation result of the paper is the following.

Theorem 5.4. *For any non-null stationary ergodic process with summable continuity rate and uniformly convergent restricted continuity rate with parameters θ_1, θ_2, k_θ, and for any $\mu_n > 0$, the empirical Markov estimator of the process with the order estimated by the bounded PML Markov order estimator $\hat{k}_n = \hat{k}_{\mathrm{PML}}(X_1^n | \eta \log n)$, $\eta > 0$, with $\frac{1}{2}\log n \le \mathrm{pen}(n) \le \mathcal{O}(\sqrt{n})$ satisfies*

$$\Pr\left(\bar{d}(X_1^n, \hat{X}[\hat{k}_n]_1^n) > \frac{\beta_2}{p_{\inf}^2} \max\left\{ \bar{\gamma}\left(\left\lfloor \frac{\eta}{\theta_2} \log n \right\rfloor \right), n^{-(1-4\eta \log(|A|^4/p_{\inf}))/(4\theta_1)} \right\} + \frac{1}{n^{1/2-\mu_n}} \right)$$

$$\leq \exp\bigl(-c_4 4^{\mu_n \log n - |\log p_{\inf}|(K_n(\eta \log n, \bar{\gamma}, c\,\mathrm{pen}(n)/n) + \log\log n/\log|A|)}\bigr)$$
$$+ \exp\left(-\frac{c_5\eta^3}{\log n} n^{\eta 2\log|A|}\right) + 2^{-s_n\,\mathrm{pen}(n)},$$

if $n \geq n_0$, where $c > 0$ is an arbitrary constant, $s_n \to \infty$ and $\beta_2, c_4, c_5, n_0 > 0$ are constants depending only on the distribution of the process.

Proof. The proof including the explicit expression of the constants is in Section 8. □

Remark 5.5. If the process X is a Markov chain of order k, then the restricted continuity rate is uniformly convergent with parameters $\theta_1 = 1$, $\theta_2 > 1$ arbitrary (arbitrarily close to 1), $k_\theta = k + 1$, and if n is sufficiently large, $K_n = k$ and

$$\max\left\{\bar{\gamma}\left(\left\lfloor \frac{\eta}{\theta_2}\log n\right\rfloor\right), n^{-(1-4\eta\log(|A|^4/p_{\inf}))/(4\theta_1)}\right\} = n^{-(1-4\eta\log(|A|^4/p_{\inf}))/(4\theta_1)}.$$

An application of the Borel–Cantelli lemma in Theorem 5.4 yields the following asymptotic result.

Corollary 5.6. *For any non-null stationary ergodic process with summable continuity rate and uniformly convergent restricted continuity rate with parameters θ_1, θ_2, k_θ, the empirical Markov estimator of the process with the order estimated by the bounded PML Markov order estimator $\hat{k}_n = \hat{k}_{\mathrm{PML}}(X_1^n | r_n)$ with $\frac{1}{2}\log n \leq \mathrm{pen}(n) \leq \mathcal{O}(\sqrt{n})$ and*

$$\frac{5\log\log n}{2\log|A|} \leq r_n \leq \mathrm{o}(\log n)$$

satisfies

$$\bar{d}(X_1^n, \hat{X}[\hat{k}_n]_1^n) \leq \frac{\beta_2}{p_{\inf}^2}\max\left\{\bar{\gamma}\left(\left\lfloor\frac{r_n}{\theta_2}\right\rfloor\right), n^{-1/(4\theta_1)}\right\}$$
$$+ \frac{(\log n)^{c_6}}{\sqrt{n}} 2^{|\log p_{\inf}| K_n(r_n, \bar{\gamma}, c\,\mathrm{pen}(n)/n)}$$

eventually almost surely as $n \to +\infty$, where $c > 0$ is an arbitrary constant, and $\beta_2, c_6 > 0$ are constants depending only on the distribution of the process.

Remark 5.7. If the memory decay of the process is slow, the first term in the bound in Corollary 5.6, the bias, is essentially $\bar{\gamma}(\lfloor r_n/\theta_2 \rfloor)$, and the second term, the variance, is maximal. If the memory decay is sufficiently fast, then the rate of the estimated order $\hat{k}_n$ and the rate of K_n are smaller, therefore the variance term is smaller, while the bias term is smaller as well. The result, however, shows the optimality of the PML Markov order estimator in the sense that it selects an order which is small enough to allow the variance to decrease but large enough to keep the bias below a polynomial threshold.

6. Empirical entropies

In this section, we consider the problem of simultaneous convergence of empirical entropies of orders in an increasing set, and prove the following theorem that formulates Theorem 4.1 with explicit constants.

Theorem 6.1. *For any weakly non-null and α-summable stationary ergodic process, for any $0<\varepsilon<1/2$*

$$\begin{aligned}&\Pr\biggl(\max_{1\le k\le(\varepsilon\log n)/(4\log|A|)}|\hat{H}_k(X_1^n)-H_k|>\frac{1}{n^{1/2-\varepsilon}}\biggr)\\&\quad\le 6\mathrm{e}^{1/\mathrm{e}}\exp\biggl(-\frac{7\alpha_0\varepsilon^3}{32\mathrm{e}(\alpha+\alpha_0)}\frac{n^{\varepsilon/2}}{\log n}+\frac{\varepsilon}{4}\log n\biggr)\end{aligned}$$

and

$$\begin{aligned}&\Pr\biggl(\max_{0\le k\le(\varepsilon\log n)/(4\log|A|)}|\hat{h}_k(X_1^n)-h_k|>\frac{1}{n^{1/2-\varepsilon}}\biggr)\\&\quad\le 12\mathrm{e}^{1/\mathrm{e}}\exp\biggl(-\frac{7\alpha_0\varepsilon^3}{256\mathrm{e}(\alpha+\alpha_0)}\frac{n^{\varepsilon/2}}{\log n}+\frac{\varepsilon}{4}\log n\biggr).\end{aligned}$$

First, we show the following bounds.

Proposition 6.2. *For any weakly non-null and α-summable stationary ergodic process, for any $1\le m\le n$ and $u,\nu>0$,*

$$\begin{aligned}&\Pr\Bigl(\max_{1\le k\le m}|\hat{H}_k(X_1^n)-H_k|>u\Bigr)\\&\quad\le 6\mathrm{e}^{1/\mathrm{e}}|A|^m\exp\biggl(\frac{\alpha_0}{8\mathrm{e}(\alpha+\alpha_0)}\frac{-(n-m+1)u^{2(1+\nu)}}{m|A|^{2m}}\\&\qquad\qquad\times\min^2\biggl\{\biggl(\frac{\mathrm{e}}{2(1+\nu^{-1})}\biggr)^{1+\nu},\frac{u^{-\nu}\log\mathrm{e}}{2m\log|A|},\frac{u^{-\nu}}{\mathrm{e}}\biggr\}\biggr)\end{aligned}$$

and

$$\begin{aligned}&\Pr\Bigl(\max_{0\le k\le m-1}|\hat{h}_k(X_1^n)-h_k|>u\Bigr)\\&\quad\le 12\mathrm{e}^{1/\mathrm{e}}|A|^m\exp\biggl(\frac{\alpha_0}{8\mathrm{e}(\alpha+\alpha_0)}\frac{-(n-m+1)(u/2)^{2(1+\nu)}}{m|A|^{2m}}\\&\qquad\qquad\times\min^2\biggl\{\biggl(\frac{\mathrm{e}}{2(1+\nu^{-1})}\biggr)^{1+\nu},\frac{(u/2)^{-\nu}\log\mathrm{e}}{2m\log|A|},\frac{(u/2)^{-\nu}}{\mathrm{e}}\biggr\}\biggr).\end{aligned}$$

Proof. Fix $1 \le k \le m$. Applying Lemma A.1 in the Appendix to the distributions $P_k = \{P(a_1^k), a_1^k \in A^k\}$ and $\hat{P}_k = \{\hat{P}(a_1^k), a_1^k \in A^k\}$,

$$|\hat{H}_k(X_1^n) - H_k| \le \frac{1}{\log \mathrm{e}}[k \log |A| - \log d_{\mathrm{TV}}(\hat{P}_k, P_k)] d_{\mathrm{TV}}(\hat{P}_k, P_k), \tag{6.1}$$

if $d_{\mathrm{TV}}(\hat{P}_k, P_k) \le 1/\mathrm{e}$. For any $\nu > 0$, the right of (6.1) can be written as

$$\begin{aligned} &\frac{k \log |A|}{\log \mathrm{e}} d_{\mathrm{TV}}(\hat{P}_k, P_k) \\ &\quad + \frac{1+\nu}{\nu \log \mathrm{e}} d_{\mathrm{TV}}^{1/(1+\nu)}(\hat{P}_k, P_k)[-d_{\mathrm{TV}}^{\nu/(1+\nu)}(\hat{P}_k, P_k) \log d_{\mathrm{TV}}^{\nu/(1+\nu)}(\hat{P}_k, P_k)] \\ &\le \frac{k \log |A|}{\log \mathrm{e}} d_{\mathrm{TV}}(\hat{P}_k, P_k) + \frac{1}{\mathrm{e}} \frac{1+\nu}{\nu} d_{\mathrm{TV}}^{1/(1+\nu)}(\hat{P}_k, P_k), \end{aligned} \tag{6.2}$$

where we used the bound $-x \log x \le \mathrm{e}^{-1} \log \mathrm{e}$, $x \ge 0$.

By [21], for any string $a_1^k \in A^k$ and $t > 0$,

$$\Pr(|N_n(a_1^k) - (n-k+1)P(a_1^k)| > t) \le \mathrm{e}^{1/\mathrm{e}} \exp\left(\frac{-c_\alpha t^2}{k(n-k+1)}\right), \tag{6.3}$$

where

$$c_\alpha = \frac{\alpha_0}{8\mathrm{e}(\alpha + \alpha_0)}$$

is positive for any weakly non-null and α-summable stationary ergodic process. (6.3) implies that

$$\begin{aligned} \Pr(d_{\mathrm{TV}}(\hat{P}_k, P_k) > t) &\le \Pr\left(\max_{a_1^k \in A^k} |\hat{P}(a_1^k) - P(a_1^k)| > \frac{t}{|A|^k}\right) \\ &\le \mathrm{e}^{1/\mathrm{e}} |A|^k \exp\left(\frac{-c_\alpha (n-k+1) t^2}{k |A|^{2k}}\right). \end{aligned} \tag{6.4}$$

Applying (6.4) to (6.2),

$$\begin{aligned} &\Pr(|\hat{H}_k(X_1^n) - H_k| > u) \\ &\le \Pr\left(\frac{k \log |A|}{\log \mathrm{e}} d_{\mathrm{TV}}(\hat{P}_k, P_k) + \frac{1}{\mathrm{e}} \frac{1+\nu}{\nu} d_{\mathrm{TV}}^{1/(1+\nu)}(\hat{P}_k, P_k) > u\right) \\ &\quad + \Pr(d_{\mathrm{TV}}(\hat{P}_k, P_k) > 1/\mathrm{e}) \\ &\le \Pr\left(d_{\mathrm{TV}}(\hat{P}_k, P_k) > \frac{u \log \mathrm{e}}{2k \log |A|}\right) + \Pr\left(d_{\mathrm{TV}}^{1/(1+\nu)}(\hat{P}_k, P_k) > \frac{\nu \mathrm{e} u}{2(1+\nu)}\right) \\ &\quad + \Pr(d_{\mathrm{TV}}(\hat{P}_k, P_k) > 1/\mathrm{e}) \end{aligned}$$

$$\le 3\mathrm{e}^{1/\mathrm{e}}|A|^k \exp\biggl(\frac{-c_\alpha(n-k+1)u^{2(1+\nu)}}{k|A|^{2k}} \times \min{}^2\biggl\{\biggl(\frac{\mathrm{e}}{2(1+\nu^{-1})}\biggr)^{1+\nu}, \frac{u^{-\nu}\log\mathrm{e}}{2k\log|A|}, \frac{u^{-\nu}}{\mathrm{e}}\biggr\}\biggr).$$

This completes the proof of the first claimed bound as

$$\begin{aligned}
&\Pr\Bigl(\max_{1\le k\le m}|\hat H_k(X_1^n)-H_k|>u\Bigr)\\
&\quad\le \sum_{1\le k\le m}\Pr(|\hat H_k(X_1^n)-H_k|>u)\\
&\quad\le 3\mathrm{e}^{1/\mathrm{e}}\Bigl(\sum_{1\le k\le m}|A|^k\Bigr)\\
&\qquad\times\exp\biggl(\frac{-c_\alpha(n-m+1)u^{2(1+\nu)}}{m|A|^{2m}}\min{}^2\biggl\{\biggl(\frac{\mathrm{e}}{2(1+\nu^{-1})}\biggr)^{1+\nu}, \frac{u^{-\nu}\log\mathrm{e}}{2m\log|A|}, \frac{u^{-\nu}}{\mathrm{e}}\biggr\}\biggr).
\end{aligned}$$

The second claimed bound follows using $\hat h_0(X_1^n)-h_0=\hat H_1(X_1^n)-H_1$ and

$$|\hat h_k(X_1^n)-h_k|\le|\hat H_{k+1}(X_1^n)-H_{k+1}|+|\hat H_k(X_1^n)-H_k|,\qquad k\ge1,$$

as

$$\begin{aligned}
&\Pr\Bigl(\max_{0\le k\le m-1}|\hat h_k(X_1^n)-h_k|>u\Bigr)\\
&\quad\le\Pr\Bigl(\max_{1\le k\le m}|\hat H_k(X_1^n)-H_k|>\frac{u}{2}\Bigr)+\Pr\Bigl(\max_{1\le k\le m-1}|\hat H_k(X_1^n)-H_k|>\frac{u}{2}\Bigr)\\
&\quad\le 2\Pr\Bigl(\max_{1\le k\le m}|\hat H_k(X_1^n)-H_k|>\frac{u}{2}\Bigr).
\end{aligned}$$

□

Now, the theorem follows from the proposition with special settings.

Proof of Theorem 6.1. We use Proposition 6.2 setting $u=n^{-1/2+\varepsilon}$, $\nu=\varepsilon$, and $m=\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor$. Then, in the exponent of the first inequality of the proposition,

$$\begin{aligned}
\frac{u^{2(1+\nu)}}{|A|^{2m}}&>n^{\varepsilon/2-1+2\varepsilon^2},\\
\frac{n-m+1}{m}&>n\frac{7}{\log n},\\
\min\biggl\{\biggl(\frac{\mathrm{e}}{2(1+\nu^{-1})}\biggr)^{1+\nu}, \frac{u^{-\nu}\log\mathrm{e}}{2m\log|A|}, \frac{u^{-\nu}}{\mathrm{e}}\biggr\}&>n^{-\varepsilon^2}\biggl(\frac{2\varepsilon}{3}\biggr)^{3/2}>n^{-\varepsilon^2}\frac{\varepsilon^{3/2}}{2},
\end{aligned}$$

where we used that $0 < \varepsilon < 1/2$. This gives the lower bound

$$-\frac{7\alpha_0\varepsilon^3}{32\mathrm{e}(\alpha+\alpha_0)}\frac{n^{\varepsilon/2}}{\log n}$$

on the exponent and completes the proof of the first claimed bound. The second claimed bound follows similarly from the second inequality of the proposition with the same settings. □

7. Divergence bounds proofs

In this section, we consider the divergence of the PML, NML and KT Markov order estimators and prove Theorems 4.5, 4.7 and 4.11.

Proof of Theorem 4.5. By [21], any weakly non-null and α-summable process is ϕ-mixing with a coefficient related to $\alpha_0 > 0$ and $\alpha < +\infty$. Namely, there exists a sequence ρ_i, $i \in \mathbb{N}$, satisfying

$$\sum_{i=0}^{\infty}\rho_i \le 1 + \frac{2\alpha}{\alpha_0},$$

such that for each k, m, l and each $a_1^k \in A^k$, $b_1^m \in A^m$, with $P(b_1^m) > 0$,

$$|\Pr(X_{m+l+1}^{m+l+k} = a_1^k | X_1^m = b_1^m) - P(a_1^k)| \le \sum_{i=l}^{l+k-1}\rho_i.$$

This implies that for any $d \ge 1$

$$\begin{aligned}
&\Pr(X_{m+l+1}^{m+l+k} = a_1^k | X_1^m = b_1^m)\\
&\quad\le \Pr(X_{m+l+id} = a_{id}, 1 \le i \le \lfloor k/d \rfloor | X_1^m = b_1^m)\\
&\quad= \prod_{i=1}^{\lfloor k/d \rfloor}\Pr(X_{m+l+id} = a_{id} | X_{m+l+jd} = a_{jd}, 1 \le j < i, X_1^m = b_1^m)\\
&\quad\le \prod_{i=1}^{\lfloor k/d \rfloor}(P(a_{id}) + \rho_{d-1})\\
&\quad\le \Big(\max_{a\in A}P(a) + \rho_{d-1}\Big)^{\lfloor k/d \rfloor}.
\end{aligned}$$

Since $\max_{a\in A}P(a) < 1$ and $\rho_d \to 0$, $\max_{a\in A}P(a) + \rho_{d-1} < 1$ for sufficiently large d. Then

$$\max_{l,a_1^k,b_1^m}\Pr(X_{m+l+1}^{m+l+k} = a_1^k | X_1^m = b_1^m) \le 2^{\lambda_1 - \lambda_2 k}$$

holds with $\lambda_1 = -\log(\max_{a\in A} P(a)+\rho_{d-1}) > 0$ and $\lambda_2 = -\log(\max_{a\in A} P(a)+\rho_{d-1})^{1/d} > 0$. Thus, for any k,

$$\begin{aligned}
&\Pr(N_n(a_1^k) \geq 2 \text{ for some } a_1^k) \\
&\quad = \Pr(X_i^{i+k-1} = X_j^{j+k-1} \text{ for some } 1 \leq i < j \leq n-k+1) \\
&\quad \leq \sum_{1\leq i<j\leq n-k+1} \Pr(X_i^{i+k-1} = X_j^{j+k-1}) \qquad (7.1) \\
&\quad = \sum_{1\leq i<j\leq n-k+1} \mathbb{E}\{\Pr(X_j^{j+k-1} = X_i^{i+k-1} | X_1^{j-1})\} \\
&\quad \leq n^2 2^{\lambda_1 - \lambda_2 k}.
\end{aligned}$$

For any information criterion IC, we can write

$$\begin{aligned}
&\{\hat{k}_{\mathrm{IC}}(X_1^n) > k_n\} \\
&\quad \subseteq \{\mathrm{IC}_{X_1^n}(m) < \mathrm{IC}_{X_1^n}(k_n) \text{ for some } m > k_n\} \\
&\quad \subseteq \{\mathrm{IC}_{X_1^n}(m) < \mathrm{IC}_{X_1^n}(k_n) \text{ for some } m > k_n\} \cap \{N_n(a_1^{k_n}) \leq 1 \text{ for all } a_1^{k_n}\} \qquad (7.2) \\
&\qquad \cup \{N_n(a_1^{k_n}) \geq 2 \text{ for some } a_1^{k_n}\}.
\end{aligned}$$

Here, $N_n(a_1^{k_n}) \leq 1$ for all $a_1^{k_n} \in A^{k_n}$ implies that $N_n(a_1^m) \leq 1$ for all $a_1^m \in A^m$ for all $m \geq k_n$, which further implies that for all $m > k_n$ (i) $\hat{h}_m(X_1^n) = 0$ and therefore $\mathrm{PML}_{X_1^n}(m) = (|A|-1)|A|^m \,\mathrm{pen}(n)$ and $\mathrm{NML}_{X_1^n}(m) = \Sigma_{n,m}$ and (ii) $\mathrm{KT}_{X_1^n}(m) = |A|^{-n}$. Then all the three information criteria do not depend on the sample and are non-decreasing in m. Hence, in (7.2)

$$\{\mathrm{IC}_{X_1^n}(m) < \mathrm{IC}_{X_1^n}(k_n) \text{ for some } m > k_n\} \cap \{N_n(a_1^{k_n}) \leq 1 \text{ for all } a_1^{k_n}\}$$

is an empty set. Thus, (7.2) gives

$$\Pr(\hat{k}_{\mathrm{IC}}(X_1^n) > k_n) \leq \Pr(N_n(a_1^{k_n}) \geq 2 \text{ for some } a_1^{k_n})$$

and using (7.1) completes the proof. □

To prove Theorem 4.7, first we show the following bounds.

Proposition 7.1. *For any weakly non-null and α-summable stationary ergodic process with $h_k - \bar{H} \leq \delta 2^{-\zeta k}$ for some $\delta, \zeta > 0$, if*

$$\frac{4\log|A|}{\zeta} \leq \varepsilon < \frac{1}{2},$$

(i) *the PML Markov order estimator* $\hat{k}_{\mathrm{PML}}(X_1^n)$ *satisfies that*

$$\Pr(\hat{k}_{\mathrm{PML}}(X_1^n) < k_n) \le 12\mathrm{e}^{1/\mathrm{e}} \exp\left(-\frac{7\alpha_0\varepsilon^3}{256\mathrm{e}(\alpha+\alpha_0)} \frac{n^{\varepsilon/2}}{\log n} + \frac{\varepsilon}{4}\log n\right),$$

if $n \ge (\delta 2^{\zeta})^2$, *where*

$$k_n = \min\left\{k \ge 0 : h_k - \bar{H} < \frac{4\max(\sqrt{n}, (|A|-1)\,\mathrm{pen}(n))}{n^{1-\varepsilon}}\right\};$$

(ii) *the Markov order estimator* $\hat{k}_{\mathrm{IC}}(X_1^n)$, *where IC is either NML or KT, satisfies that*

$$\Pr(\hat{k}_{\mathrm{IC}}(X_1^n) < k_n) \le 12\mathrm{e}^{1/\mathrm{e}} \exp\left(-\frac{7\alpha_0\varepsilon^3}{256\mathrm{e}(\alpha+\alpha_0)} \frac{n^{\varepsilon/2}}{\log n} + \frac{\varepsilon}{4}\log n\right),$$

if $n \ge \max^2\{\sqrt{24}(\log^2 \mathrm{e})(|A|-1)^2, 2C_{\mathrm{KT}}, \delta 2^{\zeta}\}$, *where*

$$k_n = \min\left\{k \ge 0 : h_k - \bar{H} < \frac{4}{n^{1/2-\varepsilon}}\right\}.$$

Remark 7.2. For Markov chains of order k, in Proposition 7.1 $k_n = k$ if n is sufficiently large.

Proof of Proposition 7.1. Let $0 < \varepsilon < 1/2$ be arbitrary and

$$B_n\left(\frac{\varepsilon\log n}{4\log|A|}\right) = \left\{\max_{0\le k\le(\varepsilon\log n)/(4\log|A|)} |\hat{h}_k(X_1^n) - h_k| \le \frac{1}{n^{1/2-\varepsilon}}\right\}. \tag{7.3}$$

For any information criterion IC, we can write for any $k_n \le \frac{\varepsilon\log n}{4\log|A|}$

$$\begin{aligned}
&\{\hat{k}_{\mathrm{IC}}(X_1^n) < k_n\} \\
&\quad\subseteq \left\{\mathrm{IC}_{X_1^n}(m) \le \mathrm{IC}_{X_1^n}\left(\left\lfloor\frac{\varepsilon\log n}{4\log|A|}\right\rfloor\right) \text{ for some } m < k_n\right\} \\
&\quad\subseteq \left(\left\{\mathrm{IC}_{X_1^n}(m) \le \mathrm{IC}_{X_1^n}\left(\left\lfloor\frac{\varepsilon\log n}{4\log|A|}\right\rfloor\right) \text{ for some } m < k_n\right\} \cap B_n\left(\frac{\varepsilon\log n}{4\log|A|}\right)\right) \\
&\qquad\cup \overline{B_n\left(\frac{\varepsilon\log n}{4\log|A|}\right)}.
\end{aligned} \tag{7.4}$$

(i) If IC = PML, by the definition of the PML information criterion, see Definition 3.3,

$$\left\{\mathrm{PML}_{X_1^n}(m) \le \mathrm{PML}_{X_1^n}\left(\left\lfloor\frac{\varepsilon\log n}{4\log|A|}\right\rfloor\right) \text{ for some } m < k_n\right\} \cap B_n\left(\frac{\varepsilon\log n}{4\log|A|}\right)$$

$$
\begin{aligned}
&\subseteq \Bigg\{ (n-m)\hat{h}_m(X_1^n) - \left(n - \left\lfloor \frac{\varepsilon \log n}{4\log|A|} \right\rfloor\right) \hat{h}_{\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor}(X_1^n) \\
&\qquad \le (|A|-1)(|A|^{\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor} - |A|^m)\,\mathrm{pen}(n) \text{ for some } m < k_n \Bigg\} \cap B_n\left(\frac{\varepsilon\log n}{4\log|A|}\right) \\
&\subseteq \Bigg\{ \hat{h}_m(X_1^n) - \hat{h}_{\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor}(X_1^n) \\
&\qquad \le (|A|-1)|A|^{\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor} \frac{\mathrm{pen}(n)}{n - \lfloor(\varepsilon\log n)/(4\log|A|)\rfloor} \text{ for some } m < k_n \Bigg\} \\
&\qquad \cap B_n\left(\frac{\varepsilon\log n}{4\log|A|}\right) \\
&\subseteq \Bigg\{ h_m - h_{\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor} \le \frac{(|A|-1)|A|^{(\varepsilon\log n)/(4\log|A|)}\,\mathrm{pen}(n)}{n - (\varepsilon\log n)/(4\log|A|)} + \frac{2}{n^{1/2-\varepsilon}} \\
&\qquad \text{for some } m < k_n \Bigg\}.
\end{aligned}
\tag{7.5}
$$

Since for any $0 < \varepsilon < 1/2$

$$
\frac{|A|^{(\varepsilon\log n)/(4\log|A|)}}{n - (\varepsilon\log n)/(4\log|A|)} < \frac{1}{n^{1-\varepsilon}}, \tag{7.6}
$$

we have

$$
\frac{(|A|-1)|A|^{(\varepsilon\log n)/(4\log|A|)}\,\mathrm{pen}(n)}{n - (\varepsilon\log n)/(4\log|A|)} + \frac{2}{n^{1/2-\varepsilon}} < \frac{3\max(\sqrt{n}, (|A|-1)\,\mathrm{pen}(n))}{n^{1-\varepsilon}}. \tag{7.7}
$$

Now, let ε and k_n be as in the claim of the proposition. Using the conditions $h_k - \bar{H} \le \delta 2^{-\zeta k}$ and $\varepsilon \ge (4\log|A|)/\zeta$,

$$
h_{\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor} - \bar{H} \le \delta \exp\left\{-\zeta\left(\frac{\varepsilon\log n}{4\log|A|} - 1\right)\right\} \le \frac{1}{\sqrt{n}} \qquad \text{if } n \ge (\delta 2^{\zeta})^2. \tag{7.8}
$$

Thus, if $n \ge (\delta 2^{\zeta})^2$, it follows that $k_n \le \frac{\varepsilon\log n}{4\log|A|}$, and for any $m < k_n$

$$
\begin{aligned}
h_m - h_{\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor} &\ge (h_{k_n-1} - \bar{H}) - (h_{\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor} - \bar{H}) \\
&\ge (h_{k_n-1} - \bar{H}) - \frac{1}{\sqrt{n}} \ge \frac{3\max(\sqrt{n}, (|A|-1)\,\mathrm{pen}(n))}{n^{1-\varepsilon}},
\end{aligned}
\tag{7.9}
$$

where we used that h_k is non-increasing. Comparing (7.9) to (7.7), the right of (7.5) is an empty set, and (7.4) yields

$$
\Pr(\hat{k}_{\mathrm{PML}}(X_1^n) < k_n) \le \Pr\left(\overline{B_n\left(\frac{\varepsilon\log n}{4\log|A|}\right)}\right) \le 12\mathrm{e}^{1/\mathrm{e}} \exp\left(-\frac{7\alpha_0\varepsilon^3}{256\mathrm{e}(\alpha+\alpha_0)} \frac{n^{\varepsilon/2}}{\log n} + \frac{\varepsilon}{4}\log n\right),
$$

if $n \ge (\delta 2^{\zeta})^2$, according to Theorem 6.1.

(ii) If IC = NML, by the definition of the NML information criterion, see Definition 3.5,

$$\begin{aligned}
&\left\{\mathrm{NML}_{X_1^n}(m) \le \mathrm{NML}_{X_1^n}\left(\left\lfloor \frac{\varepsilon \log n}{4\log|A|}\right\rfloor\right) \text{ for some } m < k_n\right\} \cap B_n\left(\frac{\varepsilon\log n}{4\log|A|}\right)\\
&\quad\subseteq \left\{(n-m)\hat{h}_m(X_1^n) - \left(n - \left\lfloor \frac{\varepsilon \log n}{4\log|A|}\right\rfloor\right)\hat{h}_{\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor}(X_1^n)\right.\\
&\qquad\left. \le \log\Sigma\left(n, \left\lfloor \frac{\varepsilon \log n}{4\log|A|}\right\rfloor\right) - \log\Sigma(n,m) \text{ for some } m < k_n\right\} \cap B_n\left(\frac{\varepsilon\log n}{4\log|A|}\right)\\
&\quad\subseteq \left\{\hat{h}_m(X_1^n) - \hat{h}_{\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor}(X_1^n) < \frac{\log\Sigma(n, \lfloor(\varepsilon\log n)/(4\log|A|)\rfloor)}{n - \lfloor(\varepsilon\log n)/(4\log|A|)\rfloor}\right.\\
&\qquad\left. \text{for some } m < k_n\right\}\\
&\qquad\cap B_n\left(\frac{\varepsilon\log n}{4\log|A|}\right)\\
&\quad\subseteq \left\{h_m - h_{\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor} < \frac{\log\Sigma(n, \lfloor(\varepsilon\log n)/(4\log|A|)\rfloor)}{n - (\varepsilon\log n)/(4\log|A|)} + \frac{2}{n^{1/2-\varepsilon}}\right.\\
&\qquad\left. \text{for some } m < k_n\right\},
\end{aligned} \tag{7.10}$$

where in the second relation we used that $\Sigma(n,m) > 1$ for any $m \ge 0$. By Lemma A.2 in the Appendix,

$$\mathrm{ML}_k(X_1^n) \le P_{\mathrm{KT},k}(X_1^n)\exp\left(C_{\mathrm{KT}}|A|^k + \frac{|A|-1}{2}|A|^k\log\frac{n}{|A|^k}\right)$$

that gives the upper bound

$$\log\Sigma(n,k) \le C_{\mathrm{KT}}|A|^k + \frac{|A|-1}{2}|A|^k\log\frac{n}{|A|^k}. \tag{7.11}$$

Using (7.11) and (7.6),

$$\begin{aligned}
&\frac{\log\Sigma(n, \lfloor(\varepsilon\log n)/(4\log|A|)\rfloor)}{n - (\varepsilon\log n)/(4\log|A|)}\\
&\quad< \left(C_{\mathrm{KT}} + \frac{|A|-1}{2}\log\frac{n}{|A|^{\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor}}\right)\frac{1}{n^{1-\varepsilon}}\\
&\quad< \left(C_{\mathrm{KT}} + \frac{|A|-1}{2}\log n\right)\frac{1}{n^{1-\varepsilon}}.
\end{aligned}$$

Using $\mathrm{e}^x \geq x^2/2 + x^4/4!$, $x \geq 0$, it follows that $(|A|-1)\log n \leq \sqrt{n}$ if $n \geq 24(\log^4 \mathrm{e})(|A|-1)^4$, which implies that

$$C_{\mathrm{KT}} + \frac{|A|-1}{2}\log n \leq \sqrt{n} \qquad \text{if } n \geq \max\{24(\log^4 \mathrm{e})(|A|-1)^4, 4C_{\mathrm{KT}}^2\}.$$

Thus, the expression in (7.10) can be bounded as

$$\begin{aligned} &\frac{\log \Sigma(n, \lfloor(\varepsilon \log n)/(4\log|A|)\rfloor)}{n - (\varepsilon \log n)/(4\log|A|)} + \frac{2}{n^{1/2-\varepsilon}} \\ &\quad < \frac{3}{n^{1/2-\varepsilon}} \qquad \text{if } n \geq \max\{24(\log^4 \mathrm{e})(|A|-1)^4, 4C_{\mathrm{KT}}^2\}. \end{aligned} \tag{7.12}$$

Now, let ε and k_n be as in the claim of the proposition. Then the conditions $h_k - \bar{H} \leq \delta 2^{-\zeta k}$ and $\varepsilon \geq (4\log|A|)/\zeta$ imply (7.8), thus, if $n \geq (\delta 2^{\zeta})^2$, it follows that $k_n \leq \frac{\varepsilon \log n}{4\log|A|}$, and for any $m < k_n$

$$\begin{aligned} h_m - h_{\lfloor(\varepsilon \log n)/(4\log|A|)\rfloor} &\geq (h_{k_n-1} - \bar{H}) - (h_{\lfloor(\varepsilon \log n)/(4\log|A|)\rfloor} - \bar{H}) \\ &\geq (h_{k_n-1} - \bar{H}) - \frac{1}{\sqrt{n}} \\ &\geq \frac{3}{n^{1/2-\varepsilon}}, \end{aligned} \tag{7.13}$$

where we used that h_k is non-increasing. Comparing (7.13) to (7.12), the right of (7.10) is an empty set, and (7.4) yields

$$\Pr(\hat{k}_{\mathrm{NML}}(X_1^n) < k_n) \leq \Pr\left(\overline{B_n\left(\frac{\varepsilon \log n}{4\log|A|}\right)}\right) \leq 12\mathrm{e}^{1/\mathrm{e}} \exp\left(-\frac{7\alpha_0 \varepsilon^3}{256\mathrm{e}(\alpha+\alpha_0)}\frac{n^{\varepsilon/2}}{\log n} + \frac{\varepsilon}{4}\log n\right),$$

if $n \geq \max\{24(\log^4 \mathrm{e})(|A|-1)^4, 4C_{\mathrm{KT}}^4, (\delta 2^{\zeta})^2\}$, according to Theorem 6.1.

(iii) If IC = KT, by the definition of the KT information criterion, see Definition 3.7, and using that $P_{\mathrm{KT},m}(X_1^n) \leq \mathrm{ML}_m(X_1^n)$ for any $0 \leq m < n$,

$$\begin{aligned} &\left\{\mathrm{KT}_{X_1^n}(m) \leq \mathrm{KT}_{X_1^n}\left(\left\lfloor\frac{\varepsilon \log n}{4\log|A|}\right\rfloor\right) \text{ for some } m < k_n\right\} \cap B_n\left(\frac{\varepsilon \log n}{4\log|A|}\right) \\ &\quad \subseteq \left\{(n-m)\hat{h}_m(X_1^n) - \left(n - \left\lfloor\frac{\varepsilon \log n}{4\log|A|}\right\rfloor\right)\hat{h}_{\lfloor(\varepsilon \log n)/(4\log|A|)\rfloor}(X_1^n)\right. \\ &\qquad \left. \leq \log \mathrm{ML}_{\lfloor(\varepsilon \log n)/(4\log|A|)\rfloor}(X_1^n) - \log P_{\mathrm{KT},\lfloor(\varepsilon \log n)/(4\log|A|)\rfloor}(X_1^n) \text{ for some } m < k_n\right\} \\ &\qquad \cap B_n\left(\frac{\varepsilon \log n}{4\log|A|}\right) \\ &\quad \subseteq \left\{\hat{h}_m(X_1^n) - \hat{h}_{\lfloor(\varepsilon \log n)/(4\log|A|)\rfloor}(X_1^n)\right. \end{aligned}$$

$$\leq \frac{\log \mathrm{ML}_{\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor}(X_1^n) - \log P_{\mathrm{KT},\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor}(X_1^n)}{n - \lfloor(\varepsilon\log n)/(4\log|A|)\rfloor} \tag{7.14}$$

$$\text{for some } m < k_n \Big\}$$

$$\cap\, B_n\left(\frac{\varepsilon\log n}{4\log|A|}\right)$$

$$\subseteq \Big\{ h_m - h_{\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor}$$

$$\leq \frac{\log \mathrm{ML}_{\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor}(X_1^n) - \log P_{\mathrm{KT},\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor}(X_1^n)}{n - (\varepsilon\log n)/(4\log|A|)} + \frac{2}{n^{1/2-\varepsilon}}$$

$$\text{for some } m < k_n \Big\}.$$

By Lemma A.2 in the Appendix,

$$\log \mathrm{ML}_{\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor}(X_1^n) - \log P_{\mathrm{KT},\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor}(X_1^n)$$
$$\leq C_{\mathrm{KT}}|A|^{(\varepsilon\log n)/(4\log|A|)} + \frac{|A|-1}{2}|A|^{(\varepsilon\log n)/(4\log|A|)}\log\frac{n}{|A|^{\lfloor(\varepsilon\log n)/(4\log|A|)\rfloor}},$$

and the proof continues in the same way as in the NML case (ii). □

Now, we are ready to prove Theorem 4.7. We prove the following theorem that formulates Theorem 4.7 with explicit constants.

Theorem 7.3. *For any weakly non-null stationary ergodic process with continuity rates* $\bar{\gamma}(k) \leq \delta_1 2^{-\zeta_1 k}$ *and* $\underline{\gamma}(k) \geq \delta_2 2^{-\zeta_2 k}$ *for some* $\zeta_1, \zeta_2, \delta_1, \delta_2 > 0$ *(*$\zeta_2 \geq \zeta_1$*), if*

$$\frac{6\log|A|}{\zeta_1} \leq \varepsilon < \frac{1}{2},$$

(i) *the PML Markov order estimator* $\hat{k}_{\mathrm{PML}}(X_1^n)$ *satisfies that*

$$\Pr(\hat{k}_{\mathrm{PML}}(X_1^n) \leq k_n) \leq 12\mathrm{e}^{1/\mathrm{e}}\exp\left(-\frac{7\alpha_0\varepsilon^3}{256\mathrm{e}(\alpha+\alpha_0)}\frac{n^{\varepsilon/2}}{\log n} + \frac{\varepsilon}{4}\log n\right),$$

if $n \geq (36\delta_1^{4/3}2^{(4\zeta_1)/3}\log^2|A|)/(\log^2\mathrm{e})$*, where*

$$k_n = \frac{1}{2\zeta_2}\left(2\log\delta_2 - 3 + \left(\frac{1}{2}-\varepsilon\right)\log n - \log\max\left\{1, (|A|-1)\frac{\mathrm{pen}(n)}{\sqrt{n}}\right\}\right);$$

(ii) *the Markov order estimator* $\hat{k}_{\mathrm{IC}}(X_1^n)$*, where IC is either NML or KT, satisfies that*

$$\Pr(\hat{k}_{\mathrm{IC}}(X_1^n) \le k_n) \le 12\mathrm{e}^{1/\mathrm{e}} \exp\left(-\frac{7\alpha_0\varepsilon^3}{256\mathrm{e}(\alpha+\alpha_0)}\frac{n^{\varepsilon/2}}{\log n} + \frac{\varepsilon}{4}\log n\right),$$

if $n \ge \max^2\{\sqrt{24}(\log^2 \mathrm{e})(|A|-1)^2, 2C_{\mathrm{KT}}, (6\delta_1^{2/3}2^{(2\zeta_1)/3}\log|A|)/(\log \mathrm{e})\}$*, where*

$$k_n = \frac{1}{2\zeta_2}\left(2\log\delta_2 - 3 + \left(\frac{1}{2}-\varepsilon\right)\log n\right).$$

[Here, C_{KT} *is the constant in the well-known bound of* $\log \mathrm{ML}_k(X_1^n) - \log P_{\mathrm{KT},k}(X_1^n)$*, see Lemma A.2 in the Appendix.]*

Proof. By Remark 2.2, $\sum_{k=0}^{+\infty}\bar{\gamma}(k) \le \sum_{k=0}^{+\infty}\delta_1 2^{-\zeta_1 k} < +\infty$ implies the α-summability. The deviation of the conditional entropies from the entropy rate will also be controlled by the continuity rates of the process, and Proposition 7.1 will yield the claim of the theorem.

First, for any $k \le m$,

$$\begin{aligned}
&h_k - h_m \\
&= \sum_{a\in A}\sum_{a_{m-k+1}^m \in A^k}\left(-P(a_{m-k+1}^m a)\log\frac{P(a_{m-k+1}^m a)}{P(a_{m-k+1}^m)}\right. \\
&\qquad\qquad \left. - \sum_{a_1^{m-k}\in A^{m-k}} -P(a_1^m a)\log\frac{P(a_1^m a)}{P(a_1^m)}\right) \qquad (7.15)\\
&= \sum_{a\in A}\sum_{a_{m-k+1}^m \in A^k}\left(-P(a_{m-k+1}^m)\sum_{a_1^{m-k}\in A^{m-k}}\frac{P(a_1^m)}{P(a_{m-k+1}^m)}\left(\frac{P(a_{m-k+1}^m a)}{P(a_{m-k+1}^m)}\log\frac{P(a_{m-k+1}^m a)}{P(a_{m-k+1}^m)}\right)\right. \\
&\qquad\qquad \left. - P(a_{m-k+1}^m)\sum_{a_1^{m-k}\in A^{m-k}} -\frac{P(a_1^m)}{P(a_{m-k+1}^m)}\left(\frac{P(a_1^m a)}{P(a_1^m)}\log\frac{P(a_1^m a)}{P(a_1^m)}\right)\right) \\
&= \sum_{a_{m-k+1}^m \in A^k} -P(a_{m-k+1}^m)\sum_{a_1^{m-k}\in A^{m-k}}\frac{P(a_1^m)}{P(a_{m-k+1}^m)} \\
&\quad \times \sum_{a\in A}\left(\frac{P(a_{m-k+1}^m a)}{P(a_{m-k+1}^m)}\log\frac{P(a_{m-k+1}^m a)}{P(a_{m-k+1}^m)} - \frac{P(a_1^m a)}{P(a_1^m)}\log\frac{P(a_1^m a)}{P(a_1^m)}\right).
\end{aligned}$$

On the right of (7.15), the difference of entropies of the conditional distributions $\{P(a|a_{m-k+1}^m), a\in A\}$ and $\{P(a|a_1^m), a\in A\}$ appears. By Remark 2.1, the total variation of these conditional distributions can be upper bounded as

$$d_{\mathrm{TV}}(P(\cdot|a_{m-k+1}^m), P(\cdot|a_1^m)) = \sum_{a\in A}|P(a|a_{m-k+1}^m) - P(a|a_1^m)| \le \bar{\gamma}(k).$$

Hence, applying Lemma A.1 in the Appendix it follows, similar to the bound (6.1) and (6.2) in the proof of Proposition 6.2, that

$$\begin{aligned}
&\left|\sum_{a\in A} P(a|a_{m-k+1}^m)\log P(a|a_{m-k+1}^m) - \sum_{a\in A} P(a|a_1^m)\log P(a|a_1^m)\right| \\
&\quad\le \frac{\log|A|}{\log \mathrm{e}} d_{\mathrm{TV}}(P(\cdot|a_{m-k+1}^m), P(\cdot|a_1^m)) + \frac{1}{\mathrm{e}}\frac{1+\nu}{\nu} d_{\mathrm{TV}}^{1/(1+\nu)}(P(\cdot|a_{m-k+1}^m), P(\cdot|a_1^m)) \\
&\quad\le \frac{\log|A|}{\log \mathrm{e}} \bar{\gamma}(k) + \frac{1}{\mathrm{e}}\frac{1+\nu}{\nu} \bar{\gamma}(k)^{1/(1+\nu)} \\
&\quad\le \frac{2\log|A|}{\log \mathrm{e}} \frac{1+\nu}{\nu} \bar{\gamma}(k)^{1/(1+\nu)}
\end{aligned} \tag{7.16}$$

for any $\nu > 0$, if $\bar{\gamma}(k) \le 1/\mathrm{e}$. Setting $\nu = 1/2$, combining (7.16) with (7.15) and taking $m \to +\infty$ yield the bound

$$h_k - \bar{H} \le \frac{6\log|A|}{\log \mathrm{e}} \bar{\gamma}(k)^{2/3}, \tag{7.17}$$

if $\bar{\gamma}(k) \le 1/\mathrm{e}$. Since $h_k - \bar{H} \le h_k \le \log|A|$, the bound (7.17) is trivial if $\bar{\gamma}(k) > 1/\mathrm{e}$. Hence, using the assumption $\bar{\gamma}(k) \le \delta_1 2^{-\zeta_1 k}$ of the theorem,

$$h_k - \bar{H} \le \frac{6\log|A|}{\log \mathrm{e}} \delta_1^{2/3} 2^{-2\zeta_1 k/3}, \tag{7.18}$$

and the assumption $h_k - \bar{H} \le \delta 2^{-\zeta k}$ of Proposition 7.1 is satisfied with

$$\delta = \frac{6\log|A|}{\log \mathrm{e}} \delta_1^{2/3} \quad \text{and} \quad \zeta = \frac{2\zeta_1}{3}.$$

Thus, the constraint $\varepsilon \ge (4\log|A|)/\zeta$ in Proposition 7.1 becomes $\varepsilon \ge (6\log|A|)/\zeta_1$, and $n \ge (\delta 2^{\zeta})^2$ becomes

$$n \ge \frac{36\log^2|A|}{\log^2 \mathrm{e}} \delta_1^{4/3} 2^{(4\zeta_1)/3}.$$

Next, for any $k < +\infty$,

$$\begin{aligned}
&h_k - \bar{H} \\
&\quad= \sum_{x_{-k}^{-1}\in A^k} \sum_{a\in A} -P(x_{-k}^{-1}a)\log P(a|x_{-k}^{-1}) \\
&\qquad + \int_{A^\infty} \sum_{a\in A} P(a|x_{-\infty}^{-1})\log P(a|x_{-\infty}^{-1})\,\mathrm{d}P(x_{-\infty}^{-1})
\end{aligned} \tag{7.19}$$

$$= \int_{A^\infty} \sum_{a\in A} P(a|x_{-\infty}^{-1}) \log \frac{P(a|x_{-\infty}^{-1})}{P(a|x_{-k}^{-1})} \,\mathrm{d}P(x_{-\infty}^{-1})$$
$$= \int_{A^\infty} D(P(\cdot|x_{-\infty}^{-1}) \| P(\cdot|x_{-k}^{-1})) \,\mathrm{d}P(x_{-\infty}^{-1}),$$

where $D(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence. Using Pinsker's inequality [8, 10], (7.19) can be lower bounded by

$$\int_{A^\infty} \frac{1}{2} \Bigg(\sum_{a\in A} |P(a|x_{-\infty}^{-1}) - P(a|x_{-k}^{-1})| \Bigg)^2 \mathrm{d}P(x_{-\infty}^{-1}) \geq \frac{1}{2} \underline{\gamma}(k)^2 \geq \delta_2^2 2^{-2\zeta_2 k - 1}, \tag{7.20}$$

where in the last inequality we used the assumption $\underline{\gamma}(k) \geq \delta_2 2^{-\zeta_2 k}$ of the theorem. Hence, in case (i)

$$\begin{aligned}
&\min\left\{ k \geq 0 : h_k - \bar{H} < \frac{4\max(\sqrt{n}, (|A|-1)\operatorname{pen}(n))}{n^{1-\varepsilon}} \right\} \\
&\quad \geq \min\left\{ k \geq 0 : \delta_2^2 2^{-2\zeta_2 k - 1} < \frac{4\max(\sqrt{n}, (|A|-1)\operatorname{pen}(n))}{n^{1-\varepsilon}} \right\} \\
&\quad = \min\left\{ k \geq 0 : k > \frac{1}{2\zeta_2} (2\log\delta_2 - 3 + (1-\varepsilon)\log n - \log\max(\sqrt{n}, (|A|-1)\operatorname{pen}(n))) \right\} \\
&\quad = 1 + \left\lfloor \frac{1}{2\zeta_2} \left(2\log\delta_2 - 3 + \left(\frac{1}{2} - \varepsilon\right)\log n - \log\max\left\{ 1, (|A|-1)\frac{\operatorname{pen}(n)}{\sqrt{n}} \right\} \right) \right\rfloor,
\end{aligned}$$

while in case (ii)

$$\begin{aligned}
&\min\left\{ k \geq 0 : h_k - \bar{H} < \frac{4}{n^{1/2-\varepsilon}} \right\} \\
&\quad \geq \min\left\{ k \geq 0 : \delta_2^2 2^{-2\zeta_2 k - 1} < \frac{4}{n^{1/2-\varepsilon}} \right\} \\
&\quad = \min\left\{ k \geq 0 : k > \frac{1}{2\zeta_2} \left(2\log\delta_2 - 3 + \left(\frac{1}{2} - \varepsilon\right)\log n \right) \right\} \\
&\quad = 1 + \left\lfloor \frac{1}{2\zeta_2} \left(2\log\delta_2 - 3 + \left(\frac{1}{2} - \varepsilon\right)\log n \right) \right\rfloor,
\end{aligned}$$

and the proof is completed. □

Finally, we prove the following proposition that directly implies Theorem 4.11.

Proposition 7.4. *For any weakly non-null stationary ergodic process with continuity rate* $\bar{\gamma}(k) \leq \delta 2^{-\zeta k}$, $\zeta, \delta > 0$, *and for any* $\xi > 0$, *if* $\varepsilon > 0$ *is so small and* $\zeta > 0$ *is so large*

that

$$\frac{1}{2}+\varepsilon<\kappa<1-\frac{\varepsilon}{4}$$

and

$$\frac{6\log|A|}{\zeta}\le\frac{\varepsilon}{1-\kappa}<2\xi,$$

the PML Markov order estimator $\hat{k}_{\mathrm{PML}}(X_1^n)$ *with* $\mathrm{pen}(n)=n^\kappa$ *satisfies that*

$$\Pr\left(\left|\frac{\hat{k}_{\mathrm{PML}}(X_1^n)}{k_{\mathrm{PML},n}}-1\right|>\xi\right)\le\exp\left(-\frac{c_2'\varepsilon^3}{\log n}n^{\varepsilon/2}\right),$$

if n *is sufficiently large, where* $c_2'>0$ *is a constant depending only on the distribution of the process.*

Proof. The proof of Theorem 7.3 begins with the observation that the summability of the continuity rate implies the α-summability. Hence, the conditions of Theorem 6.1 are satisfied now. Moreover, according to (7.18), $\bar{\gamma}(k)\le\delta 2^{-\zeta k}$ also implies that

$$h_k-\bar{H}\le\frac{6\log|A|}{\log\mathrm{e}}\delta^{2/3}2^{-2\zeta k/3}. \tag{7.21}$$

Set ξ, ε and κ as in the conditions of the proposition, and define a sequence $k_n\in\mathbb{N}$ such that for sufficiently large n

$$\text{(i)}\quad h_{\lfloor(1-\xi/2)k_n\rfloor}-h_{k_n}\ge|A|n^{-1+\kappa+\varepsilon/4},$$

$$\text{(ii)}\quad h_{k_n}-\bar{H}\le\frac{1}{2}(|A|-1)^2n^{-1+\kappa},$$

$$\text{(iii)}\quad k_n\le\frac{\varepsilon\log n}{4\log|A|}.$$

Due to (7.21), such a sequence exists. Since $h_k-\bar{H}$ is non-negative decreasing, it is sufficient to show this when $h_k-\bar{H}=\frac{6\log|A|}{\log\mathrm{e}}\delta^{2/3}2^{-2\zeta k/3}$. Then, writing k_n in the form $k_n=\nu\log n$,

$$h_{\lfloor(1-\xi/2)k_n\rfloor}-h_{k_n}\ge\frac{1}{2}\frac{6\log|A|}{\log\mathrm{e}}\delta^{2/3}n^{-2\zeta(1-\xi/2)\nu/3}\quad\text{and}\quad h_{k_n}-\bar{H}=\frac{6\log|A|}{\log\mathrm{e}}\delta^{2/3}n^{-2\zeta\nu/3},$$

if n is sufficiently large, that implies (i) and (ii) if

$$1-\kappa<\frac{2\zeta\nu}{3}<\left(1-\kappa-\frac{\varepsilon}{4}\right)\frac{1}{1-\xi/2}.$$

Such $\nu>0$ exists because it follows from the condition $\varepsilon/(1-\kappa)<2\xi$ that $1-\kappa<(1-\kappa-\varepsilon/4)/(1-\xi/2)$. Moreover, the condition $\varepsilon/(1-\kappa)\ge(6\log|A|)/\zeta$ implies $\nu\le\varepsilon/(4\log|A|)$ satisfying (iii).

First, recall the definition of $B_n(\frac{\varepsilon \log n}{4 \log |A|})$ in (7.3) in the proof of Proposition 7.1. Similar to (7.4) and (7.5), we can write that

$$\{\hat{k}_{\mathrm{PML}}(X_1^n) < (1-\xi/2)k_n\} \cap B_n\left(\frac{\varepsilon \log n}{4 \log |A|}\right) \tag{7.22}$$

$$\subseteq \{\mathrm{PML}_{X_1^n}(m) \le \mathrm{PML}_{X_1^n}(k_n) \text{ for some } m < (1-\xi/2)k_n\} \cap B_n\left(\frac{\varepsilon \log n}{4 \log |A|}\right)$$

$$= \{\mathrm{PML}_{o,n}(m) + (n-m)(\hat{h}_m(X_1^n) - h_m) \le \mathrm{PML}_{o,n}(k_n) + (n-k_n)(\hat{h}_{k_n}(X_1^n) - h_{k_n})$$
$$\text{for some } m < (1-\xi/2)k_n\} \cap B_n\left(\frac{\varepsilon \log n}{4 \log |A|}\right)$$

$$\subseteq \left\{\mathrm{PML}_{o,n}(m) - \frac{2n}{n^{1/2-\varepsilon}} \le \mathrm{PML}_{o,n}(k_n) \text{ for some } m < (1-\xi/2)k_n\right\} \tag{7.23}$$

$$\subseteq \left\{(n-m)h_m - (n-k_n)h_{k_n} \le (|A|-1)(|A|^{k_n} - |A|^m)\,\mathrm{pen}(n) + \frac{2n}{n^{1/2-\varepsilon}}\right.$$
$$\left.\text{for some } m < (1-\xi/2)k_n\right\}$$

$$\subseteq \left\{h_m - h_{k_n} \le \frac{(|A|-1)|A|^{(\varepsilon \log n)/(4 \log |A|)}\,\mathrm{pen}(n)}{n - (\varepsilon \log n)/(4 \log |A|)} + \frac{2}{n^{1/2-\varepsilon}} \text{ for some } m < (1-\xi/2)k_n\right\}$$

$$\subseteq \{h_{\lfloor (1-\xi/2)k_n \rfloor} - h_{k_n} < |A| n^{-1+\kappa+\varepsilon/4}\} \tag{7.24}$$

that is empty set by (i), if n is large enough and $k_n \le \frac{\varepsilon \log n}{4 \log |A|}$. The latter is satisfied because of (iii). On the other hand,

$$\{\hat{k}_{\mathrm{PML}}(X_1^n) > (1+\xi/2)k_n\} \cap B_n\left(\frac{\varepsilon \log n}{4 \log |A|}\right) \cap \left\{\hat{k}_{\mathrm{PML}}(X_1^n) \le \frac{\varepsilon \log n}{4 \log |A|}\right\} \tag{7.25}$$

$$\subseteq \left\{\mathrm{PML}_{X_1^n}(m) < \mathrm{PML}_{X_1^n}(k_n) \text{ for some } (1+\xi/2)k_n < m \le \frac{\varepsilon \log n}{4 \log |A|}\right\}$$
$$\cap B_n\left(\frac{\varepsilon \log n}{4 \log |A|}\right)$$

$$\subseteq \left\{\mathrm{PML}_{o,n}(m) - \frac{2n}{n^{1/2-\varepsilon}} < \mathrm{PML}_{o,n}(k_n) \text{ for some } m > (1+\xi/2)k_n\right\} \tag{7.26}$$

$$\subseteq \left\{(|A|-1)(|A|^m - |A|^{k_n})\,\mathrm{pen}(n) - \frac{2n}{n^{1/2-\varepsilon}} < (n-k_n)h_{k_n} - (n-m)h_m\right.$$
$$\left.\text{for some } m > (1+\xi/2)k_n\right\}$$

$$\subseteq \Bigg\{ (|A|-1)(|A|^m - |A|^{k_n}) \frac{\mathrm{pen}(n)}{n} - \frac{2}{n^{1/2-\varepsilon}} < h_{k_n} - \left(1 - \frac{m}{n}\right) h_m \quad \text{for some } m > (1+\xi/2)k_n \Bigg\}$$

$$\subseteq \Bigg\{ (|A|-1)(|A|^m - |A|^{k_n}) \frac{\mathrm{pen}(n)}{n} - \frac{2}{n^{1/2-\varepsilon}} - \frac{m}{n}\bar{H} < (h_{k_n} - \bar{H}) - \left(1 - \frac{m}{n}\right)(h_{k_n} - \bar{H}) \quad \text{for some } m > (1+\xi/2)k_n \Bigg\}$$

$$\subseteq \left\{ h_{k_n} - \bar{H} > \frac{(|A|-1)^2}{2} n^{-1+\kappa} \right\} \tag{7.27}$$

that is empty set by (ii), if n is large enough.

Observe that

$$\frac{1+\xi/2}{1+\xi} k_n \le k_{\mathrm{PML},n} \le \frac{1-\xi/2}{1-\xi} k_n, \tag{7.28}$$

if n is sufficiently large. Indeed, on indirect way the following sequence of implications can be written

$$\begin{aligned} k_{\mathrm{PML},n} < \frac{1+\xi/2}{1+\xi} k_n \quad &\Rightarrow \quad k_{\mathrm{PML},n} < 1 - \frac{\xi/2}{1+\xi} k_n \quad \Rightarrow \quad k_{\mathrm{PML},n} < (1-\xi/2)k_n \\ &\Rightarrow \quad \mathrm{PML}_{o,n}(m) < \mathrm{PML}_{o,n}(k_n) \qquad \text{for some } m < (1-\xi/2)k_n \\ &\Rightarrow \quad \mathrm{PML}_{o,n}(m) - \frac{2n}{n^{1/2-\varepsilon}} < \mathrm{PML}_{o,n}(k_n) \\ &\qquad\quad \text{for some } m < (1-\xi/2)k_n \end{aligned}$$

that does not hold by (7.23) and (7.24) if n is large enough, and

$$\begin{aligned} k_{\mathrm{PML},n} > \frac{1-\xi/2}{1-\xi} k_n \quad &\Rightarrow \quad k_{\mathrm{PML},n} > 1 + \frac{\xi/2}{1-\xi} k_n \Rightarrow k_{\mathrm{PML},n} > (1+\xi/2)k_n \\ &\Rightarrow \quad \mathrm{PML}_{o,n}(m) < \mathrm{PML}_{o,n}(k_n) \qquad \text{for some } m > (1+\xi/2)k_n \\ &\Rightarrow \quad \mathrm{PML}_{o,n}(m) - \frac{2n}{n^{1/2-\varepsilon}} < \mathrm{PML}_{o,n}(k_n) \\ &\qquad\quad \text{for some } m > (1+\xi/2)k_n \end{aligned}$$

that does not hold either by (7.26) and (7.27) if n is large enough.

Finally, using (7.28), we get

$$\begin{aligned}
&\Pr\left(\left|\frac{\hat{k}_{\mathrm{PML}}(X_1^n)}{k_{\mathrm{PML},n}}-1\right|>\xi\right)\\
&\quad=\Pr(\hat{k}_{\mathrm{PML}}(X_1^n)<(1-\xi)k_{\mathrm{PML},n})+\Pr(\hat{k}_{\mathrm{PML}}(X_1^n)>(1+\xi)k_{\mathrm{PML},n})\\
&\quad\le\Pr(\hat{k}_{\mathrm{PML}}(X_1^n)<(1-\xi/2)k_n)+\Pr(\hat{k}_{\mathrm{PML}}(X_1^n)>(1+\xi/2)k_n)\\
&\quad\le\Pr\left(\{\hat{k}_{\mathrm{PML}}(X_1^n)<(1-\xi/2)k_n\}\cap B_n\left(\frac{\varepsilon\log n}{4\log|A|}\right)\right)\\
&\qquad+\Pr\left(\{\hat{k}_{\mathrm{PML}}(X_1^n)>(1+\xi/2)k_n\}\cap B_n\left(\frac{\varepsilon\log n}{4\log|A|}\right)\cap\left\{\hat{k}_{\mathrm{PML}}(X_1^n)\le\frac{\varepsilon\log n}{4\log|A|}\right\}\right)\\
&\qquad+2\Pr\left(\overline{B_n\left(\frac{\varepsilon\log n}{4\log|A|}\right)}\right)+\Pr\left(\hat{k}_{\mathrm{PML}}(X_1^n)>\frac{\varepsilon\log n}{4\log|A|}\right),
\end{aligned}$$

where the first two terms are zero if n is large enough by (7.22)–(7.24) and (7.25)–(7.27). Using Proposition 8.3 with $r_n=n-1$, $k_n=\lfloor\frac{\varepsilon\log n}{4\log|A|}\rfloor$ and $m_n=\lfloor\frac{\varepsilon\log n}{6\log|A|}\rfloor$,

$$\Pr\left(\hat{k}_{\mathrm{PML}}(X_1^n)>\frac{\varepsilon\log n}{4\log|A|}\right)\le\exp(-\mathcal{O}(n^{\kappa+\varepsilon/4})),$$

because

$$n\bar{\gamma}(m_n)\le n\delta 2^{\zeta}\exp\left(-\zeta\frac{\varepsilon\log n}{6\log|A|}\right)=\delta 2^{\zeta}n^{1-\zeta\varepsilon/(6\log|A|)},$$

but $1-\zeta\varepsilon/(6\log|A|)<\kappa+\varepsilon/4$ according to the condition $\varepsilon/(1-\kappa)\ge(6\log|A|)/\zeta$. Then the claim of the proposition follows from Theorem 6.1. □

8. Process estimation proofs

In this section, we consider the estimation of stationary ergodic processes by finite memory processes. First, define

$$\beta_1=\frac{1}{\prod_{j=1}^{+\infty}(1-2\bar{\gamma}(j))}$$

and

$$\beta_2=\sup_{k\ge1}2|A|\frac{1-(1-2|A|\bar{\gamma}(k))^k}{k\bar{\gamma}(k)\prod_{j=1}^{+\infty}(1-2|A|\bar{\gamma}(j))^2}.$$

Clearly, if $\sum_{k=1}^{\infty}\bar{\gamma}(k)<+\infty$, then $\beta_1,\beta_2<+\infty$.

Now we prove the following theorem that formulates Theorem 5.4 with explicit constants.

Theorem 8.1. *For any non-null stationary ergodic process with summable continuity rate and uniformly convergent restricted continuity rate with parameters θ_1, θ_2, k_θ, for any $\mu_n > 0$, the empirical Markov estimator of the process with the order estimated by the bounded PML Markov order estimator $\hat{k}_{\mathrm{PML}}(X_1^n|\eta\log n)$, $\eta > 0$, with penalty function $\mathrm{pen}(n) \le \mathcal{O}(\sqrt{n})$ satisfies*

$$\begin{aligned}
&\Pr\biggl(\bar{d}(X_1^n, \hat{X}[\hat{k}_{\mathrm{PML}}(X_1^n|\eta\log n)]_1^n) > \frac{\beta_2}{p_{\mathrm{inf}}^2} g_n + \frac{1}{n^{1/2-\mu_n}}\biggr) \\
&\quad \le 2\mathrm{e}^{1/\mathrm{e}}|A|^{K_n+h_n+2} \exp\biggl\{ -\frac{p_{\mathrm{inf}}^2}{16\mathrm{e}|A|^3(\alpha+p_{\mathrm{inf}})(\beta_1+1)^2} \frac{(n-K_n-h_n)}{(1+K_n+h_n)n} \\
&\qquad\qquad \times 4^{-(K_n+h_n)|\log p_{\mathrm{inf}}|}\biggl[4^{\mu_n\log n} - \frac{(K_n+h_n)|\log p_{\mathrm{inf}}|(\beta_1+1)^2}{2}\biggr]\biggr\} \\
&\qquad + 12\mathrm{e}^{1/\mathrm{e}} \exp\biggl(-\frac{7\alpha_0(\log|A|)^3\eta^3}{4\mathrm{e}(\alpha+\alpha_0)} \frac{n^{\eta 2\log|A|}}{\log n} + (\eta\log|A|)\log n\biggr) \\
&\qquad + \exp\biggl(-(|A|-1)|A|^{K_n+h_n+1} \\
&\qquad\qquad \times \mathrm{pen}(n)\biggl[1 - \frac{1}{|A|^{1+h_n}} - \frac{1}{2\,\mathrm{pen}(n)}(\log n - (K_n+h_n)\log|A|)\biggr] \\
&\qquad\qquad + \frac{c\,\mathrm{pen}(n)}{p_{\mathrm{inf}}/\log\mathrm{e}} + |A|^{K_n+h_n+1}C_{\mathrm{KT}} + \log(\eta\log n)\biggr),
\end{aligned}$$

if n is so large that

$$\min\biggl\{\biggl\lfloor\frac{\eta}{\theta_2}\log n\biggr\rfloor, k \ge 0 : \bar{\gamma}(k) < \biggl(\frac{6\max(\sqrt{n}, (|A|-1)\,\mathrm{pen}(n))}{p_{\mathrm{inf}} n^{1-\eta\log(|A|^4/p_{\mathrm{inf}})}}\biggr)^{1/(2\theta_1)}\biggr\} \ge k_\theta, \qquad (8.1)$$

where

$$\begin{aligned}
g_n &= \max\biggl\{\bar{\gamma}\biggl(\biggl\lfloor\frac{\eta}{\theta_2}\log n\biggr\rfloor\biggr), \biggl(\frac{6\max(1, (|A|-1)(\mathrm{pen}(n))/\sqrt{n})}{p_{\mathrm{inf}} n^{1/2-\eta\log(|A|^4/p_{\mathrm{inf}})}}\biggr)^{1/(2\theta_1)}\biggr\}, \\
K_n &= K_n\biggl(r_n, \bar{\gamma}, \frac{c}{n}\,\mathrm{pen}(n)\biggr),
\end{aligned}$$

and $c > 0$ is an arbitrary constant and $h_n \in \mathbb{N}$ is an arbitrary sequence.

The proof is based on the following two propositions.

Proposition 8.2. *For any non-null and α-summable stationary ergodic process with uniformly convergent restricted continuity rate with parameters θ_1, θ_2, k_θ,*

(i) *the bounded PML Markov order estimator* $\hat{k}_{\mathrm{PML}}(X_1^n|\eta\log n)$ *with penalty function* $\mathrm{pen}(n)\le\mathcal{O}(\sqrt{n})$ *satisfies that*

$$\Pr(\hat{k}_{\mathrm{PML}}(X_1^n|\eta\log n)<k_n)\le 12\mathrm{e}^{1/\mathrm{e}}\exp\biggl(-\frac{7\alpha_0(\log|A|)^3\eta^3}{4\mathrm{e}(\alpha+\alpha_0)}\frac{n^{\eta 2\log|A|}}{\log n}+(\eta\log|A|)\log n\biggr),$$

if n *is so large that* $k_n\ge k_\theta$, *where*

$$k_n=\min\biggl\{\biggl\lfloor\frac{\eta}{\theta_2}\log n\biggr\rfloor,k\ge 0:\bar{\gamma}(k)<\biggl(\frac{6\max(\sqrt{n},(|A|-1)\,\mathrm{pen}(n))}{p_{\inf}n^{1-\eta\log(|A|^4/p_{\inf})}}\biggr)^{1/(2\theta_1)}\biggr\};$$

(ii) *the bonded Markov order estimator* $\hat{k}_{\mathrm{IC}}(X_1^n|\eta\log n)$, *where IC is either NML or KT, satisfies that*

$$\Pr(\hat{k}_{\mathrm{IC}}(X_1^n|\eta\log n)<k_n)\le 12\mathrm{e}^{1/\mathrm{e}}\exp\biggl(-\frac{7\alpha_0(\log|A|)^3\eta^3}{4\mathrm{e}(\alpha+\alpha_0)}\frac{n^{\eta 2\log|A|}}{\log n}+(\eta\log|A|)\log n\biggr),$$

if n *is so large that* $k_n\ge k_\theta$ *and* $n\ge\max^2\{\sqrt{24}(\log^2\mathrm{e})(|A|-1)^2,2C_{\mathrm{KT}}\}$, *where*

$$k_n=\min\biggl\{\biggl\lfloor\frac{\eta}{\theta_2}\log n\biggr\rfloor,k\ge 0:\bar{\gamma}(k)<\biggl(\frac{6}{p_{\inf}n^{1/2-\eta\log(|A|^4/p_{\inf})}}\biggr)^{1/(2\theta_1)}\biggr\}.$$

Proof. First, define $B_n(\eta\log n)$ similar to (7.3) in the proof of Proposition 7.1. Similar to (7.4)–(7.7), we can write for any $k_n\le(\eta/\theta_2)\log n$ that

$$\begin{aligned}&\Pr(\hat{k}_{\mathrm{PML}}(X_1^n|\eta\log n)<k_n)\\&\quad\le\Pr\biggl(h_m-h_{\lfloor\eta\log n\rfloor}<\frac{3\max(\sqrt{n},(|A|-1)\,\mathrm{pen}(n))}{n^{1-\eta 4\log|A|}}\text{ for some }m<k_n\biggr)\\&\qquad+\Pr(\overline{B_n(\eta\log n)}).\end{aligned}\tag{8.2}$$

Now, the difference $h_m-h_{\lfloor\eta\log n\rfloor}$ in (8.2) is controlled as follows. For any $m\le k$,

$$\begin{aligned}&h_m-h_k\\&\quad=\sum_{a\in A}\sum_{a_1^k\in A^k}(-P(a_1^ka)\log P(a|a_{k-m+1}^k)+P(a_1^ka)\log P(a|a_1^k))\\&\quad=\sum_{a_1^k\in A^k}P(a_1^k)\sum_{a\in A}P(a|a_1^k)\log\frac{P(a|a_1^k)}{P(a|a_{k-m+1}^k)}\\&\quad=\sum_{a_1^k\in A^k}P(a_1^k)D(P(\cdot|a_1^k)\|P(\cdot|a_{k-m+1}^k)).\end{aligned}\tag{8.3}$$

Using Pinsker's inequality [8, 10], (8.3) can be lower bounded by

$$\begin{aligned}&\sum_{a_1^k\in A^k} P(a_1^k)\frac{1}{2}\Bigg(\sum_{a\in A}|P(a|a_1^k)-P(a|a_{k-m+1}^k)|\Bigg)^2\\&\quad\geq \frac{1}{2}\bar{\gamma}(m|k)^2\min_{a_1^k\in A^k}P(a_1^k)\\&\quad\geq \frac{1}{2}\bar{\gamma}(m|k)^2 p_{\mathrm{inf}}^k.\end{aligned}\tag{8.4}$$

Using (8.4) and the assumption $\bar{\gamma}(k)^{\theta_1}\leq\bar{\gamma}(k|\lceil\theta_2 k\rceil)$ if $k\geq k_\theta$ $(\theta_1\geq 1,\theta_2>1)$, it follows that

$$h_k-h_{\lceil\theta_2 k\rceil}\geq\tfrac{1}{2}\bar{\gamma}(k|\lceil\theta_2 k\rceil)^2 p_{\mathrm{inf}}^{\lceil\theta_2 k\rceil}\geq\tfrac{1}{2}\bar{\gamma}(k)^{2\theta_1}p_{\mathrm{inf}}^{\theta_2 k+1}\qquad\text{if } k\geq k_\theta.$$

Hence, we can write

$$\begin{aligned}&\min\Bigg\{k\geq k_\theta : h_k-h_{\lceil\theta_2 k\rceil}<\frac{3\max(\sqrt{n},(|A|-1)\,\mathrm{pen}(n))}{n^{1-\eta 4\log|A|}}\Bigg\}\\&\quad\geq\min\Bigg\{k\geq k_\theta:\bar{\gamma}(k)<\Bigg(\frac{6\max(\sqrt{n},(|A|-1)\,\mathrm{pen}(n))}{n^{1-\eta 4\log|A|}}2^{-(\theta_2 k+1)\log p_{\mathrm{inf}}}\Bigg)^{1/(2\theta_1)}\Bigg\}\\&\quad\geq\min\Bigg\{k\geq k_\theta:\bar{\gamma}(k)<\Bigg(\frac{6\max(\sqrt{n},(|A|-1)\,\mathrm{pen}(n))}{p_{\mathrm{inf}}n^{1-\eta\log(|A|^4/p_{\mathrm{inf}})}}\Bigg)^{1/(2\theta_1)}\Bigg\}.\end{aligned}\tag{8.5}$$

Let k_n be as in the claim of the proposition and suppose that $k_n\geq k_\theta$. Then, since h_k is non-increasing, for any $m<k_n\leq(\eta/\theta_2)\log n$

$$h_m-h_{\lfloor\eta\log n\rfloor}\geq h_{k_n-1}-h_{\lceil\theta_2(k_n-1)\rceil}\geq\frac{3\max(\sqrt{n},(|A|-1)\,\mathrm{pen}(n))}{n^{1-\eta 4\log|A|}}.\tag{8.6}$$

Applying (8.6) to (8.2), the first term on the right in (8.2) equals zero, therefore

$$\begin{aligned}\Pr(\hat{k}_{\mathrm{PML}}(X_1^n)<k_n)&\leq\Pr(\overline{B_n(\eta\log n)})\\&\leq 12\mathrm{e}^{1/\mathrm{e}}\exp\Bigg(-\frac{7\alpha_0(\log|A|)^3\eta^3}{4\mathrm{e}(\alpha+\alpha_0)}\frac{n^{\eta 2\log|A|}}{\log n}+(\eta\log|A|)\log n\Bigg)\end{aligned}$$

by Theorem 6.1 with $\varepsilon=\eta 4\log|A|$.

In cases $\mathrm{IC}=\mathrm{NML}$ and $\mathrm{IC}=\mathrm{KT}$, the proofs deviate from the above similar to as (ii) and (iii) deviate from (i) in the proof of Proposition 7.1. Now, instead of (7.12) we have

$$\begin{aligned}&\frac{\log\Sigma(n,\lfloor\eta\log n\rfloor)}{n-\eta\log n}+\frac{2}{n^{1/2-\eta 4\log|A|}}\\&\quad<\frac{3}{n^{1/2-\eta 4\log|A|}}\qquad\text{if } n\geq\max\{24(\log^4\mathrm{e})(|A|-1)^4,4C_{\mathrm{KT}}^2\}.\end{aligned}$$

□

Proposition 8.3. *For any non-null stationary ergodic process, the bounded PML Markov order estimator* $\hat{k}_{\mathrm{PML}}(X_1^n|r_n)$ *satisfies that*

$$\begin{aligned}&\Pr(\hat{k}_{\mathrm{PML}}(X_1^n|r_n) > k_n)\\ &\quad\le \exp\Bigg(\log(r_n - k_n) + \frac{(n-m_n)\bar{\gamma}(m_n)}{p_{\inf}/\log \mathrm{e}} + (|A|-1)|A|^{m_n}\,\mathrm{pen}(n)\\ &\qquad + |A|^{k_n+1}\left[C_{\mathrm{KT}} + \frac{|A|-1}{2}\log\frac{n}{|A|^{k_n+1}} - (|A|-1)\,\mathrm{pen}(n)\right]\Bigg)\end{aligned}$$

for any $0 \le m_n \le k_n \le r_n \le n$.

Proof. For any $m \ge 0$,

$$P(x_1^n) = P(x_1^m)\prod_{i=m+1}^{n} P(x_i|x_1^{i-1}) \le \left(\prod_{i=m+1}^{n} P(x_i|x_{i-m}^{i-1})\right)\prod_{i=m+1}^{n}\frac{P(x_i|x_1^{i-1})}{P(x_i|x_{i-m}^{i-1})}. \tag{8.7}$$

Using $P(x_i|x_1^{i-1}) \le P(x_i|x_{i-m}^{i-1}) + \bar{\gamma}(m)$ and $P(x_i|x_{i-m}^{i-1}) \ge p_{\inf}$, (8.7) can be upper bounded by

$$\left(\prod_{i=m+1}^{n} P(x_i|x_{i-m}^{i-1})\right)\left(1+\frac{\bar{\gamma}(m)}{p_{\inf}}\right)^{n-m} \le \mathrm{ML}_m(x_1^n)\left(1+\frac{\bar{\gamma}(m)}{p_{\inf}}\right)^{n-m}. \tag{8.8}$$

Now, let $C_{n,k} = \{\hat{k}_{\mathrm{PML}}(X_1^n|r_n) = k\}$. By the definition of the PML information criterion, see Definition 3.3, for any $0 \le m_n, k \le r_n$

$$\begin{aligned}\log \mathrm{ML}_{m_n}(X_1^n) \le \log \mathrm{ML}_k(X_1^n) &- (|A|-1)|A|^k\,\mathrm{pen}(n)\\ &+ (|A|-1)|A|^{m_n}\,\mathrm{pen}(n) \qquad \text{if } X_1^n \in C_{n,k}.\end{aligned} \tag{8.9}$$

By Lemma A.2 in the Appendix,

$$\mathrm{ML}_k(X_1^n) \le P_{\mathrm{KT},k}(X_1^n)\exp\left(C_{\mathrm{KT}}|A|^k + \frac{|A|-1}{2}|A|^k\log\frac{n}{|A|^k}\right). \tag{8.10}$$

Combining (8.8), (8.9) and (8.10),

$$\begin{aligned}P(X_1^n) \le P_{\mathrm{KT},k}(X_1^n)&\left(1+\frac{\bar{\gamma}(m_n)}{p_{\inf}}\right)^{n-m_n}\\ &\times\exp\Bigg(C_{\mathrm{KT}}|A|^k + \frac{|A|-1}{2}|A|^k\log\frac{n}{|A|^k}\\ &\qquad - (|A|-1)|A|^k\,\mathrm{pen}(n) + (|A|-1)|A|^{m_n}\,\mathrm{pen}(n)\Bigg) \qquad \text{if } X_1^n \in C_{n,k},\end{aligned}$$

that implies

$$\begin{aligned} P(C_{n,k}) &\le \left(1+\frac{\bar\gamma(m_n)}{p_{\inf}}\right)^{n-m_n} \\ &\quad\times \exp\Bigg(C_{\mathrm{KT}}|A|^k + \frac{|A|-1}{2}|A|^k \log\frac{n}{|A|^k} \\ &\qquad - (|A|-1)|A|^k \operatorname{pen}(n) + (|A|-1)|A|^{m_n}\operatorname{pen}(n)\Bigg) \qquad (8.11) \\ &\le \exp\Bigg(\frac{(n-m_n)\bar\gamma(m_n)}{p_{\inf}/\log \mathrm{e}} + (|A|-1)|A|^{m_n}\operatorname{pen}(n) \\ &\qquad + |A|^k\left[C_{\mathrm{KT}} + \frac{|A|-1}{2}\log\frac{n}{|A|^k} - (|A|-1)\operatorname{pen}(n)\right]\Bigg), \end{aligned}$$

where in the last inequality we used $\log(1+x)\le x\log \mathrm{e}$, $x\ge 0$. In the exponent of (8.11), it may be assumed that $|A|^k$ is multiplied by a negative number otherwise the bound is trivial. Then, the claim of the lemma follows from (8.11) as

$$\Pr(\hat k_{\mathrm{PML}}(X_1^n|r_n) > k_n) \le \sum_{k=k_n+1}^{r_n} P(C_{n,k}) \le (r_n-k_n)P(C_{n,k_n+1}).$$

□

Now, we are ready to prove Theorem 8.1.

Proof of Theorem 8.1. Letting

$$G_n = \{\bar\gamma(\hat k_{\mathrm{PML}}(X_1^n|\eta\log n)) \le g_n\}$$

and

$$H_n = \{\hat k_{\mathrm{PML}}(X_1^n|\eta\log n) \le k_n\},$$

write

$$\begin{aligned} &\Pr\left(\bar d(X_1^n, \hat X[\hat k_{\mathrm{PML}}(X_1^n|\eta\log n)]_1^n) > \frac{\beta_2}{p_{\inf}^2}g_n + \frac{1}{n^{1/2-\mu_n}}\right) \\ &\quad\le \Pr\left(\left\{\bar d(X_1^n, \hat X[\hat k_{\mathrm{PML}}(X_1^n|\eta\log n)]_1^n) > \frac{\beta_2}{p_{\inf}^2}g_n + \frac{1}{n^{1/2-\mu_n}}\right\}\cap G_n\cap H_n\right) \\ &\qquad + \Pr(\bar G_n) + \Pr(\bar H_n) \qquad (8.12) \\ &\quad\le \Pr\left(\left\{\bar d(X_1^n, \hat X[\hat k_{\mathrm{PML}}(X_1^n|\eta\log n)]_1^n) > \frac{\beta_2}{p_{\inf}^2}\bar\gamma(\hat k_{\mathrm{PML}}(X_1^n|\eta\log n)) + \frac{1}{n^{1/2-\mu_n}}\right\}\cap H_n\right) \\ &\qquad + \Pr(\bar G_n) + \Pr(\bar H_n). \end{aligned}$$

The three terms on the right of (8.12) is bounded as follows.

Since the process is non-null with summable continuity rate, Lemma A.3 in the Appendix with $\mu = \mu_n$, $\nu \log n = k_n$ and $k = \hat{k}_{\mathrm{PML}}(X_1^n | \eta \log n)$ gives

$$\begin{aligned}
&\Pr\left(\left\{ \bar{d}(X_1^n, \hat{X}[\hat{k}_{\mathrm{PML}}(X_1^n|\eta\log n)]_1^n) > \frac{\beta_2}{p_{\mathrm{inf}}^2} \bar{\gamma}(\hat{k}_{\mathrm{PML}}(X_1^n|\eta\log n)) + \frac{1}{n^{1/2-\mu_n}} \right\} \cap H_n \right) \\
&\quad \le 2\mathrm{e}^{1/\mathrm{e}} |A|^{k_n+2} \exp\left\{ -\frac{p_{\mathrm{inf}}^2}{16\mathrm{e}|A|^3(\alpha + p_{\mathrm{inf}})(\beta_1+1)^2} \frac{(n-k_n)4^{-k_n|\log p_{\mathrm{inf}}|}}{(1+k_n)n} \right. \\
&\qquad \left. \times \left[4^{\mu_n \log n} - \frac{k_n |\log p_{\mathrm{inf}}| (\beta_1+1)^2}{2} \right] \right\}. \qquad (8.13)
\end{aligned}$$

By Remark 2.2, the summability of the continuity rate implies the α-summability. Hence, for the non-null process with summable continuity rate and uniformly convergent restricted continuity rate with parameters θ_1, θ_2, k_θ, Proposition 8.2 implies that

$$\Pr(\bar{G}_n) \le 12\mathrm{e}^{1/\mathrm{e}} \exp\left(-\frac{7\alpha_0 (\log|A|)^3 \eta^3}{4\mathrm{e}(\alpha+\alpha_0)} \frac{n^{\eta 2 \log|A|}}{\log n} + (\eta \log|A|) \log n \right), \qquad (8.14)$$

if (8.1) holds because

$$\begin{aligned}
&\Pr(\bar{\gamma}(\hat{k}_{\mathrm{PML}}(X_1^n|\eta\log n)) \ge g_n) \\
&\quad = \Pr\left(\hat{k}_{\mathrm{PML}}(X_1^n|\eta\log n) \right. \\
&\qquad \left. \le \min\left\{ \left\lfloor \frac{\eta}{\theta_2} \log n \right\rfloor, k \ge 0 : \bar{\gamma}(k) < \left(\frac{6\max(\sqrt{n}, (|A|-1)\,\mathrm{pen}(n))}{p_{\mathrm{inf}} n^{1-\eta\log(|A|^4/p_{\mathrm{inf}})}} \right)^{1/(2\theta_1)} \right\} \right).
\end{aligned}$$

Applying Proposition 8.3 with $r_n = \eta \log n$,

$$m_n = \min\left\{ \lfloor \eta \log n \rfloor, k \ge 0 : \bar{\gamma}(k) < \frac{c\,\mathrm{pen}(n)}{n} \right\}$$

and $k_n = h_n + m_n$, it follows that

$$\begin{aligned}
\Pr(\bar{H}_n) \le \exp\Bigg(&-(|A|-1)|A|^{k_n+1} \\
&\times \mathrm{pen}(n) \left[1 - \frac{1}{|A|^{1+h_n}} - \frac{1}{2\,\mathrm{pen}(n)} (\log n - k_n \log|A|) \right] \qquad (8.15) \\
&+ \frac{c\,\mathrm{pen}(n)}{p_{\mathrm{inf}}/\log\mathrm{e}} + |A|^{k_n+1} C_{\mathrm{KT}} + \log(\eta \log n) \Bigg).
\end{aligned}$$

Finally, applying the bounds (8.13), (8.14) and (8.15) to the right of (8.12), the proof is complete. □

Appendix

Lemma A.1. *For two probability distributions P_1 and P_2 on A^k,*

$$|H(P_1) - H(P_2)| \le \frac{1}{\log \mathrm{e}}[k \log |A| - \log d_{\mathrm{TV}}(P_1, P_2)] d_{\mathrm{TV}}(P_1, P_2),$$

if $d_{\mathrm{TV}}(P_1, P_2) \le 1/\mathrm{e}$, where

$$H(P_i) = -\sum_{a_1^k \in A^k} P_i(a_1^k) \log P_i(a_1^k)$$

is the entropy of P_i, $i = 1, 2$, and

$$d_{\mathrm{TV}}(P_1, P_2) = \sum_{a_1^k \in A^k} |P_1(a_1^k) - P_2(a_1^k)|$$

is the total variation distance of P_1 and P_2.

Proof. See Lemma 3.1 of [32]. □

Lemma A.2. *There exists a constant C_{KT} depending only on $|A|$, such that for any $0 \le k < n$*

$$\log \mathrm{ML}_k(X_1^n) - \log P_{\mathrm{KT},k}(X_1^n) \le C_{\mathrm{KT}} |A|^k + \frac{|A| - 1}{2} |A|^k \log \frac{n}{|A|^k}.$$

Proof. The bound, see, for example, (27) in [9],

$$\left| \log P_{\mathrm{KT},k}(X_1^n) + k \log |A| - \log \mathrm{ML}_k(X_1^n) + \frac{|A| - 1}{2} \sum_{\substack{a_1^k \in A^k: \\ N_{n-1}(a_1^k) \ge 1}} \log N_{n-1}(a_1^k) \right|$$

$$\le C'_{\mathrm{KT}} |A|^k,$$

where C'_{KT} depends only on $|A|$, implies the claim using

$$\sum_{\substack{a_1^k \in A^k: \\ N_{n-1}(a_1^k) \ge 1}} \log N_{n-1}(a_1^k) \le |A|^k \log \frac{n}{|A|^k},$$

see Proof of Theorem 6 in [9]. □

Lemma A.3. *Let X be a non-null stationary ergodic process with summable continuity rate. Then, for any $\mu > 0$ and $k \le \nu \log n$, $\nu > 0$, the empirical k-order Markov estimator*

of the process satisfies

$$\begin{aligned}&\Pr\left\{\bar{d}(X_1^n,\hat{X}[k]_1^n) > \beta_2 p_{\inf}^{-2}\bar{\gamma}(k) + \frac{1}{n^{1/2-\mu}}\right\}\\&\quad\le 2\mathrm{e}^{1/\mathrm{e}}|A|^{2+\nu\log n}\\&\qquad\times\exp\left\{-\frac{p_{\inf}^2}{16\mathrm{e}|A|^3(\alpha+p_{\inf})(\beta_1+1)^2}\frac{(n-\nu\log n)n^{-2\nu|\log p_{\inf}|}}{(1+\nu\log n)n}\right.\\&\qquad\qquad\left.\times\left[n^{2\mu}-\frac{\nu|\log p_{\inf}|(\beta_1+1)^2\log n}{2}\right]\right\}.\end{aligned}$$

Proof. See the proof of Theorem 2 and Lemma 3 in [13]. □

Acknowledgements

The author would like to thank the referees for their comments that helped improving the presentation of the results and generalizing the consistency concept. The research of the author was supported in part by NSF Grant DMS-09-06929.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)* 267–281. Budapest: Akadémiai Kiadó. MR0483125

[2] Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413. MR1679028

[3] Barron, A., Rissanen, J. and Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory* **44** 2743–2760. MR1658898

[4] Berbee, H. (1987). Chains with infinite connections: Uniqueness and Markov representation. *Probab. Theory Related Fields* **76** 243–253. MR0906777

[5] Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)* **3** 203–268. MR1848946

[6] Bressaud, X., Fernández, R. and Galves, A. (1999). Speed of $\overline{d}$-convergence for Markov approximations of chains with complete connections. A coupling approach. *Stochastic Process. Appl.* **83** 127–138. MR1705603

[7] Comets, F., Fernández, R. and Ferrari, P.A. (2002). Processes with long memory: Regenerative construction and perfect simulation. *Ann. Appl. Probab.* **12** 921–943. MR1925446

[8] Cover, T.M. and Thomas, J.A. (2006). *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley. MR2239987

[9] Csiszár, I. (2002). Large-scale typicality of Markov sample paths and consistency of MDL order estimators. *IEEE Trans. Inform. Theory* **48** 1616–1628. MR1909476

[10] Csiszár, I. and Körner, J. (2011). *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge: Cambridge Univ. Press. MR2839250

[11] Csiszár, I. and Shields, P.C. (2000). The consistency of the BIC Markov order estimator. *Ann. Statist.* **28** 1601–1619. MR1835033

[12] Csiszár, I. and Talata, Z. (2006). Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inform. Theory* **52** 1007–1016. MR2238067

[13] Csiszár, I. and Talata, Z. (2010). On rate of convergence of statistical estimation of stationary ergodic processes. *IEEE Trans. Inform. Theory* **56** 3637–3641. MR2798525

[14] Dedecker, J. and Doukhan, P. (2003). A new covariance inequality and applications. *Stochastic Process. Appl.* **106** 63–80. MR1983043

[15] Dedecker, J. and Prieur, C. (2005). New dependence coefficients. Examples and applications to statistics. *Probab. Theory Related Fields* **132** 203–236. MR2199291

[16] Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089

[17] Duarte, D., Galves, A. and Garcia, N.L. (2006). Markov approximation and consistent estimation of unbounded probabilistic suffix trees. *Bull. Braz. Math. Soc. (N.S.)* **37** 581–592. MR2284889

[18] Fernández, R. and Galves, A. (2002). Markov approximations of chains of infinite order. *Bull. Braz. Math. Soc. (N.S.)* **33** 295–306. MR1978829

[19] Finesso, L., Liu, C.C. and Narayan, P. (1996). The optimal error exponent for Markov order estimation. *IEEE Trans. Inform. Theory* **42** 1488–1497. MR1426225

[20] Gabrielli, D., Galves, A. and Guiol, D. (2003). Fluctuations of the empirical entropies of a chain of infinite order. *Math. Phys. Electron. J.* **9** Paper 5, 17 pp. (electronic). MR2028333

[21] Galves, A. and Leonardi, F. (2008). Exponential inequalities for empirical unbounded context trees. In *In and Out of Equilibrium. 2. Progress in Probability* **60** 257–269. Basel: Birkhäuser. MR2477385

[22] Gao, J. and Gijbels, I. (2008). Bandwidth selection in nonparametric kernel testing. *J. Amer. Statist. Assoc.* **103** 1584–1594. MR2504206

[23] Krichevsky, R.E. and Trofimov, V.K. (1981). The performance of universal encoding. *IEEE Trans. Inform. Theory* **27** 199–207. MR0633417

[24] Leonardi, F. (2010). Some upper bounds for the rate of convergence of penalized likelihood context tree estimators. *Braz. J. Probab. Stat.* **24** 321–336. MR2643569

[25] Marton, K. (1998). Measure concentration for a class of random processes. *Probab. Theory Related Fields* **110** 427–439. MR1616492

[26] Ornstein, D.S. (1973). An application of ergodic theory to probability theory. *Ann. Probab.* **1** 43–58. MR0348831

[27] Ornstein, D.S. and Weiss, B. (1990). How sampling reveals a process. *Ann. Probab.* **18** 905–930. MR1062052

[28] Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry. World Scientific Series in Computer Science* **15**. Singapore: World Scientific. MR1082556

[29] Ryabko, B. and Astola, J. (2006). Universal codes as a basis for time series testing. *Stat. Methodol.* **3** 375–397. MR2252392

[30] Ryabko, B.Y. (1984). Twice-universal coding. *Probl. Inf. Transm.* **20** 173–177.

[31] Ryabko, B.Y. (1988). Prediction of random sequences and universal coding. *Probl. Inf. Transm.* **24** 87–96.

[32] Schönhuth, A. (2009). On analytic properties of entropy rate. *IEEE Trans. Inform. Theory* **55** 2119–2127. MR2729869

[33] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

[34] Shields, P.C. (1996). *The Ergodic Theory of Discrete Sample Paths. Graduate Studies in Mathematics* **13**. Providence, RI: Amer. Math. Soc. MR1400225

[35] van Handel, R. (2011). On the minimal penalty for Markov order estimation. *Probab. Theory Related Fields* **150** 709–738. MR2824872

[36] Šťar'kov, J.M. (1977). Coding of discrete sources with unknown statistics. In *Topics in Information Theory (Second Colloq., Keszthely, 1975). Colloq. Math. Soc. János Bolyai* **16** 559–574. Amsterdam: North-Holland. MR0530212

Received April 2010 and revised June 2011